The neuroscience of consciousness and its metaphysics

Whiteley, Cecily

Awarding institution:
King's College London

The Neuroscience of Consciousness and Its Metaphysics

A Neuroscience-First Approach to the Metaphysics of Mind

by

Cecily Whiteley

A Thesis Submitted for the Degree of Master of Philosophical Studies

(MPhil.Stud)

KING'S COLLEGE LONDON

September 2018

Thesis Abstract

Much of contemporary philosophical work on consciousness is predicated on the assumption that work produced as part of the neuroscience of consciousness, the now prolific and well established research field within cognitive neuroscience, is neutral - or metaphysically unilluminating - with respect to theory choice between radically divergent theories in the metaphysics of mind. The operation of this ' *neutrality assumption'* is evidenced most clearly in the methodological practices of metaphysicians and neuroscientists, who stand united in the assertion that questions about the metaphysics of consciousness and the details of its neural basis form largely independent, non-interacting research areas (Seth 2010, Howhy & Bayne 2016, Goff 2017). Left unchallenged, this has led interalia to the perceived viability and subsequent propagation of so-called naturalistic non-physicalist metaphysical accounts of consciousness, on the basis that the numerous positions falling under this heading are "broadly consistent with the picture of consciousness and the brain emerging from contemporary neuroscience" (Crane et al. 2015).

In this thesis I present a case against the neutrality assumption and the view of the metaphysics of mind it supports. This is split into three parts. In the first, I draw out the main line of argument in favour of the neutrality assumption, which is introduced in Chalmers (2002, 2010) and recently developed further in Kriegel (*forthcoming*) . With the use of case studies from recent neuroscientific research, I present two objections to this argument. Together, these aim to demonstrate that the argument depends, in various ways, on an outdated and empirically implausible view of contemporary neuroscientific research and, in particular, its *explanatory* aims and practices. In the second part, I build on this discussion and the related work it draws upon (Revonsuo 2000, Neisser 2012) to motivate an alternative methodology for the metaphysics of consciousness. On this account, neuroscientific explanation and its' ontological commitments play an indispensable role in metaphysical theorising, serving as the mutual starting constraints on an empirically adequate, naturalistic metaphysics of consciousness. In part III, I demonstrate how this methodological stance can be put to use, and put forward a number of arguments against contemporary theories in the metaphysics of mind.

Acknowledgements

I would like to thank my supervisors Matthew Soteriou, David Papineau and Matthew Parrot for their invaluable help and support throughout this project and during my MPhil.Stud programme. I would also like to thank the London Arts and Humanities Partnership for funding this research project and those people who have given me useful feedback on this material at the ASSC 22, the BPPA and research seminars at King's College London this year.

Table of Contents

Chapter 1
Introducing the Neutrality Assumption

When it comes to the contemporary study of consciousness - understood in the familiar phenomenal sense as *subjective experience;* mental states for which it is 'something it is like' to undergo for the subjects which have them - two prominent fields of study lay claim. The first - analytic philosophy of mind - is unsurprising. Rooted in a long history and tradition, contemporary philosophical work on consciousness, focusing primarily on the question of consciousness' *ontological status,* continues to construct and debate the viability of various competing metaphysical theories of consciousness. The second - neuroscience, broadly understood - is much more recent. Prompted by the growing acceptance of the view that consciousness is now a suitable topic for serious scientific investigation, the neuroscience of consciousness has subsequently emerged as a prolific and well established research field in its own right. Aiming inter alia at the provision of a comprehensive and

naturalised account of consciousness' *neurobiological basis*, the work produced as part of this research programme is extensive and ongoing.

Whilst these adjacent research programmes receive significant philosophical attention individually, the relationship between the two has received little, if any, sustained philosophical elucidation. This thesis aims to begin to rectify this, and proposes to examine the relationship between these two contemporary lines of research. Despite its absence from mainstream discussion, a default or standard conception of the relationship between the neuroscience and metaphysics of consciousness has since emerged as one of *neutrality*. That is, according to proponents of the neutrality assumption, the extensive work produced as part of the neuroscience of consciousness is *neutral* with respect to theory choice between, and construction of, competing ontological accounts of consciousness in the contemporary metaphysics of mind. In the first instance then, the aim of this thesis is to elucidate and asses the justificatory basis of this standard assumption concerning the relationship between the neuroscience and metaphysics of consciousness. In this chapter I set out the historical basis and exemplification of this neutrality claim from both a philosophical and empirical perspective, and draw out what I take to be its crucial dialectical function viz. to sustain the (robust) *methodological independence* of the metaphysics of mind from the neuroscience of consciousness. That is, I motivate the claim that the neutrality assumption concerning the relationship between the metaphysics and neuroscience of consciousness does crucial work in contemporary metaphysics of mind in virtue of justifying the *methodology* adopted in current debates, according to which metaphysicians are free to settle questions of consciousness ontological status independently from a sustained appreciation, and examination of, the details of neuroscientific work. After having set out this neutrality claim and its dialectical function in more detail, I finish with an outline of the remaining thesis, which starts from the assessment of its viability.

1 The Metaphysics of Consciousness: An Overview

The assumption that contemporary work on the neural basis of consciousness is ontologically indifferent toward the metaphysics of consciousness - the claim I call the 'neutrality assumption' - is widespread. In order to make sense of this claim and highlight what is at stake in its acceptance or rejection, the contemporary dialectic in the metaphysics of mind - in which the operation of this neutrality assumption is most evident - requires brief explication. Here the focus is on providing an informative overview of the main dialectical developments which is sufficient to introduce and make salient a number of claims about the *methodology* utilised and adopted in contemporary metaphysics of mind, which is to be discussed in detail in this thesis.

1 1 The Empirical Case for Physicalism

Contemporary dialectic in the metaphysics of consciousness, restricted here to the debate over the question of consciousness' ontological status, starts from the rise and widespread acceptance of physicalism. This can be understood (perhaps) least controversially as the view that the world's fundamental ontology is physical. That is, as the metaphysical thesis that everything is, or is some sense reducible to, the physical[1]. The rise of physicalism as the dominant metaphysical worldview within the last century, while often credited to an overzealous attitude towards science and its future successes, is in fact attributable to an empirical argument which became available only within the last hundred years following a specific empirical development viz. the scientific acceptance of the causal completeness of physics[2]. According to this *causal closure argument*, modern science from the 20th century onwards (as opposed to previous frameworks) provides strong inductive reason to think that the physical realm is causally closed, or in other words, that something like the causal closure principle - the thesis that every particular physical effect has a sufficient physical cause - is true[3].

The causal closure principle provides us with a strong argument in favour of physicalism. If true, it implies that anything capable of producing physical effects - that is, capable of figuring into causal explanations for physical events - including the conscious mental states and properties previously postulated to explain human behaviour, must be given an ontological analysis in physical terms. The conclusion of the causal closure argument applied to phenomenally conscious properties is thus the familiar disjunctive one: either phenomenal properties are

---

[1]Formulations of physicalism tend to differ along two dimensions, depending on how (i) the physical and (ii) the 'some sense reducible to' terms are defined (See Tiehen *forthcoming* and Goff 2017; chapter 1 for informative summaries). With respect to (i), *theory-based* conceptions look to define the physical in terms of the entities denoted or quantified over in current or future theories from physics (Chalmers 1996, 2012, Melnyk 1997, 2003, Witmer *forthcoming*), whereas *object-based* conceptions characterise the physical in terms of whether the entity figures in a complete account of the intrinsic nature of paradigmatic physical objects - tables, chairs and trees etc - (Jackson 1998, Stoljar 2001, 2010, Strawson 2006). In reaction to now-standard objections raised against to these conceptions (Crane and Mellor 1990, Montero 1999, 2012, Ney 2008), the most popular approach is now to adopt a *negative* conception of the physical, which attempts to define physical entities in terms of what they are not viz. non [fundamentally] mental (Spurret and Papineau 1999, Papineau 2001, Montero and Papineau 2005, Wilson 2006, Tiehen 2016) and/or entities which do not figure in value laden explanations (Goff 2017). With respect to (ii), replacing traditional *supervenience-based* conceptions of physicalism (Davidson 1970, Lewis 1983, Kim 1993) which have since been shown to be insufficient for the sort of metaphysical dependency required for physicalism (Wilson 1999, 2005, Melnyk 2003) and pre-Kripkean *a priori entailment* formulations, contemporary approaches look to characterise the dependency relation stated above in terms of the popular neo-Aristotelian relation of *Grounding* (Dasgupta 2014, Audi 2012, Schaffer 2009, Rosen 2010, Goff 2017) or, more specifically, a [subset account of] the *realisation relation* (Melnyk 2003, Gillett 2002, 2003, Endicott 2012, Wilson 2005, 2011). In this thesis I intend to be neutral between these definitions of physicalism and bring in specific formulations explicitly only when appropriate (see part III). I discuss *methodological accounts* of physicalism, which reject a purely metaphysical definition (Van Fraassen 2002, Ney 2008) in Chapter 3.

[2] Papineau (2000), (2002) and (2008) and Loewer (1995).

[3] See Papineau (2002, appendix) for a historical treatment of the inductive case for the causal closure principle. The case splits into two complementary arguments, one of which takes as its starting point the observation of successful reductive explanations of diverse phenomena in terms of *a few fundamental [physical] forces*, and the other which inductively infers the causal closure principle from the *lack of neurophysiological evidence* for causally efficacious non-physical forces in living systems. This latter argument is crucial in the current context.

physical, and thus capable of exerting their postulated causal influence, or they are *epiphenomenal* viz. non-physical properties existing in causal isolation from the physical world. Given the counter-intuitive nature of the epiphenomenalist position, and the reductio arguments brought against it which exploit the consequences the account has for the possibility of self-knowledge (Chalmers 1996, De Brigard 2014, Moore 2014) and/or moral responsibility (List & Menzies 2017, Kim 2007), the majority of philosophers did, and continue to, take the causal closure argument as decisive in favour of the physicalists' metaphysical thesis[4].

While there is broad agreement then among physicalists that the causal closure argument demands a physicalist ontology, contemporary physicalists disagree on how strict the mandate of the causal closure argument is with respect to its conclusion. Motivated by standard multiple realisation arguments against traditional so-called 'reductive' varieties of physicalism (Putnam 1967, 1975, Fodor 1974, Kitcher 1982) which seek to identify phenomenal properties with their physical supervient bases (Place 1956, Smart 1959, Lewis 1966), the majority of physicalists have since adopted a *non-reductive* stance[5]. Arguably the most popular account in the metaphysics of mind and special sciences, non-reductive physicalism is considered to be an attractive position insofar as it combines and jointly endorses the following three claims: (i) *efficacy*: mental properties, *qua* mental, can be causes and effects of other properties, (ii) *distinctness*: mental properties are ontologically irreducible and non identical to physical properties and (iii) *supervience*: mental properties nevertheless metaphysically supervene on, and thus can be given an ontological analysis ultimately in terms of, physical properties. Whilst successfully avoiding the aforementioned multiple realisation worries, non-reductive physicalists famously face problems of their own viz. accounting for mental causation. Demonstrating that the non-reductive physicalist, in attempting to bypass the problems faced by type-identity theories, inherits the causal worries which plague the dualist is the mandate of the causal exclusion argument (Kim 1998, 2005). Taking broadly the same line of argument as the causal closure argument, this has as its conclusion that the joint endorsement of (i-iii) characteristic of non-reductive physicalism is untenable, and that non-reductive physicalists must either adopt an epiphenomenalist account of phenomenal properties - thus rejecting (i) - or else, a reductive type-identity physicalist - rejecting (ii) - position[6].

---

[4] C.f Jackson (1982), Chalmers (2003), Robinson (2006, 2018), Stoljar & List (2017), Gibb (2015).

[5] Contemporary type-identity theorists include, inter alia, Levin (1991), McLaughlin (1999), Bechtel and McCauley (1999), Polger (2004), Kim [conditionally] (1998, 2005). Multiple realisation arguments against type identity theory have recently come under attack, and it remains an open question whether, and to what extent, the objection is successful in providing a compelling argument against type-identity theory in favour of a non-reductive variety of physicalism. See Polger and Shapiro, (2016). As such, in this thesis I treat the sort of sophisticated type-identity theory which has the resources to resist these standard multiple realisation objections (to be discussed in part III) as a serious ontological account of consciousness.

[6] The need to provide a compelling account of mental causation on behalf of the non-reductive physicalist is reflected in the huge amount of literature which has emerged since the exclusion argument's introduction, much of which seeks to

1 2 The Anti-Physicalist Arguments

Contemporary debate in the metaphysics of mind in the last fifty years however has been concerned primarily with the success of a different set of a priori arguments, which purport to refute *all* contemporary physicalist accounts of consciousness. Taking Chalmers' (1996) hard problem of consciousness as their mandate, the so-called *anti-physicalist arguments* - typically, the knowledge argument (Jackson 1982), the explanatory gap argument (Levine 1983, 2001), and the conceivability argument (Chalmers 1996, 2010) - claim to demonstrate that *any* physicalist metaphysics is ill equipped and ultimately fails to account for the reality of phenomenal consciousness. Their reasoning is simple. Starting from the assertion of an *epistemic* gap between phenomenal and physical descriptions - the change in Mary's epistemic situation once she has left her black and white room, the negative ideal conceivability of phenomenal zombies, the lack of a priori entailment between physical and phenomenal truths, and so on - the familiar anti-physicalist claim is that such epistemic deficiencies are indicative of, and can be used to establish, an *ontological deficiency* within the physicalists ontology. Given that a physicalist metaphysics is incompatible with the presence of such ontological deficiencies, the anti-physicalist concludes, physicalism is false. Phenomenal consciousness, the anti-physicalist maintains, involves fundamental non-physical properties [7].

Since their introduction, anti-physicalist arguments have generated a huge amount of secondary literature which, in various ways, seeks to argue for and against their viability. Two types of physicalist responses can generally be distinguished: (i) physicalists which *deny* the presence of a epistemic gap between physical and phenomenal truths (or otherwise claim that these can be easily closed), and (ii) those which *grant* that the anti-physicalist arguments succeed in establishing an epistemic gap between physical and phenomenal facts but seek to block the anti-physicalist move from this epistemic claim to the ontological conclusion required for the falsity of physicalism. The later 'a posteriori' or 'type-B' physicalist response is widely considered to be the more promising[8]. Development of an adequate type-B response on behalf of the physicalist is the mandate of the so-called Phenomenal Concept Strategy[9]. According to proponents of the phenomenal concept strategy, the epistemic gaps which motivate the anti-physicalist arguments arise not from the non-physical nature of

---

draw on contemporary metaphysical analyses of causation to block the argument's conclusion. Philosophers adopting the various so-called 'causal overdetermination' strategies include Fodor(1989), Horgan(1997, 2001), Bennett(2003, 2008), Kallestrup (2006), Shapiro and Sober (2012), Woodward (2008), Yablo(1992), List & Menzies (2016) Raatiken(2010) Zhong(2011) Loewer (2007) Pereboom(2002), Schaffer(2003), and Sider (2003) Crisp and Warfield (2001), Block(2003), Carey(2011) and Roche (2014).

[7] See also Nagel (1974), Kirk (1974), Kripke (1980), Robinson (1982, 1993), Nida-Rumelin (2007) and Goff (2011, 2017).

[8] A priori or 'type-A' physicalists however include: Ryle (1949), Dennett (1991), Dretske (1995), Nemirow (1990), Conee (1994), Jackson (2007), who usually adopt a form of analytic functionalism or behaviourism.

[9] Horgan (1984), Loar (1990, 2003), Papineau (1993, 1998, 2002), Tye (1995), Lycan (1996), Hill and McLaughlin (1997), Balog (1999, 2012), Diaz-Leon (2008, 2010), and Howell (2013).

experience itself, but rather from the special cognitive features implicated in our thinking about or *conceptualisation of* consciousness. That is, type-B physicalists maintain that our phenomenal concepts, which are available only to subjects which have had the relevant experience, have an especially intimate link to their referents - being for example, recognitional, quotational/constitutional , demonstrative or information-theoretic - which explain the puzzling aspects of our epistemic relation to our conscious states (e.g. their lack of a priori connections to physical concepts) in a manner which is consistent with physicalism. Given that the puzzling features serving as the initial motivation for the anti-physicalist position can be given a satisfactory *physical* explanation, the argument goes, the anti-physicalist arguments fail to pose a significant challenge to the physicalists position.

1 3 The Return of Two Radical Research Programmes

An extensive and complex literature exists which concerns the success of the phenomenal concept strategy as an adequate response to anti-physicalist arguments[10]. The now growing consensus however seems to be that, several layers into the secondary literature, the ontological debate has reached a stalemate. That is, there seems to be compelling - but arguably indecisive - arguments proposed on either side of the ontological divide about which there is no readily available consensus[11]. As a result of this, and the assumption that there is no further empirical means of settling the dispute, the focus of contemporary debate has changed. Within the non-physicalist camp, efforts have shifted from defense of the anti-physicalist arguments to explication of the alternative *naturalistic non-physicalist* accounts of consciousness that these arguments purport to motivate - typically, naturalistic property dualism, Russellian monism and contemporary forms of idealism[12]. Here, the

[10] Debate has focused for example, on whether the phenomenal concept strategy fails for general apriori reasons or follows from a two-dimensional account of modality (Chalmers 2007 c.f. Carruthers and Viellet 2007, Balog 2012), and/or whether the *directness* of phenomenal concepts raises problems for the physicalist who claims that the necessary a posteriori identities at issue are analogous to other cases whose falsity is similarly conceivable, yet which are presented using indirect concepts. Block & Stalnaker (1999) Chalmers and Jackson (2001), Chalmers (2002) Levine (2001, 2010). According to one contemporary diagnosis of this dialectic however, the debate between the a posteriori physicalist and her opponent can be given a simpler analysis, and ultimately comes down to whether one takes phenomenal concepts to be *revelatory,* that is, whether they reveal the essential features of their referent in virtue of our being directly acquainted with them (Goff 2017;chp5, Papineau *forthcoming*).

[11] See Balog (*in press*) for an argument for the view that whether one takes the anti-physicalist arguments to be decisive in favour of either view will depend on which (physicalist vs non-physicalist) account one starts with. "It seems like there are no principles *outside* the physicalist and anti-physicalist systems that could settle this issue. What we have here is a puzzling symmetry between the two positions. The situation seems to be at a stalemate" (18). Note, something like this seems to be the case even if one takes the revelatory diagnosis of the debate just mentioned [Balog's diagnosis is slightly different]. Here too there seems to be general acknowledgement that whether one takes an introspective faculty like acquaintance which is required for the revelatory reading of phenomenal concepts (and, by extension, a non-physicalist conclusion), to be plausible and whose naturalistically inexplicable nature is unproblematic (c.f. Papineau forthcoming) will depend on one's 'bedrock' or starting commitments (Goff 2017; chp1, chp5).

[12] Examples of property dualists include Robinson (1982), Gibb (2015), Kroedel (2015). For panpsychism and its variations: Bruntrup and Jaskolla (2017),Goff (2017), Seager (2002), Strawson (2006, 2016), and the accounts outlined in

aim is to provide the foundations of a detailed and comprehensive alternative to the physicalists' account of consciousness *consistent with a broadly naturalistic outlook*[13]. That is, whilst philosophers working within this programme take there to be compelling arguments in favour of a non-physicalist position, the view seems to be that proper assessment of the anti-physicalists dialectical position requires a more holistic treatment, including a consideration of the independent merits of the alternative theories proposed, and the capability of their proponents to overcome standard obstacles and internal difficulties pressed against them.

The final dialectical development relevant to this introduction is the emergence of a second contemporary line of research, which attempts to make progress on the hard problem of consciousness from a different angle. Moving away from acceptance of the hard problem and attempting to outline potential solutions to it, the mandate of the so-called 'meta-problem of consciousness' research programme is to provide an explanation of why we think that that consciousness poses a hard problem to begin with. That is, explain the source of our *problem reports* about the hard problem, and in doing so, "shed light on its potential solutions" [14]. Most of the past and contemporary work within this line of research has focused on the meta-problem's most radical variation, the illusion problem, whose solution is thought to support the illusionists thesis viz. that consciousness is an introspective illusion caused by the systematic mis-representation of physical-functional 'conscious' states as having qualitative phenomenal properties which they in fact lack [15]. For illusionists, solving or providing an explanation for the meta or 'illusion' problem in physical terms subsumes, and indeed dissolves, the need to provide a solution to the hard problem on which contemporary debate has fixated. At the very least, its proponents claim, the illusionist thesis ought to serve as the leading explanatory hypothesis which ought to be ruled out before more radical alternative hypotheses (that is, those developed as part of the naturalistic non-physicalist research programme) are considered.

2 The Neuroscience of Consciousness

Chalmers (2013). Contemporary idealists include Adams (2007), Kastrup (2017), Yetter-Chappell (forthcoming) and the different varieties of an idealist account described in Chalmers (forthcoming). For Russellian monism see (Goff 2017, Goff and Coleman forthcoming, Alter and Nagasawa 2015).

[13]It is thus in the emergence of the contemporary naturalistic non-physicalist research programme that the operation of the neutrality assumption is most explicitly evidenced. See the quote from Tim Crane on page 14 below.

[14]Chalmers (forthcoming) sets out the formal interdisciplinary research programme, however work on the meta-problem in various forms has been going on for a while - see Papineau (2002;chp6) and the work of illusionists discussed below.

[15] See Frankish (2016, 2012), Dennett (2016), Granzino (2013), Humphrey (2011), Clark (2000), among others. Despite being a growing but still minority position, I include the meta-problem and illusionists in my overview here for two reasons (a) to show the pervasiveness of the neutrality assumption in all sides of the contemporary debate (Frankish, for example, takes the case in favour of illusionism to be independent of what neuroscientific work *demands*) and (b) so that I can later bring in some of the methodological commitments that the (supposedly neutral) meta-problem programme commits to, which I think lends evidential support to the central methodological claim defended in this thesis.

Whilst this debate over consciousness' ontological status has waged in contemporary metaphysics and philosophy of mind, research on consciousness has taken off from a different direction. Prompted by the decline of the lingering behaviourist view in psychology and related fields that consciousness is 'taboo', and an unsuitable topic for serious empirical investigation, a new scientific research programme has emerged in the past thirty years which aims to make progress on the problem of consciousness, and develop a naturalistic account of its *neural basis*, using empirical methods. As one leading neuroscientist puts it "developing a naturalized account of the rich experiential tapestry of consciousness is now recognized as a major objective for twenty- first century science" (Seth 2010). The so-called 'new science' of consciousness - which I'll abbreviate here to the standard, but somewhat misleading, label 'the neuroscience of consciousness' - is now a large and well established interdisciplinary enterprise, comprised of thousands of researchers in neuroscience, psychology, artificial intelligence, computer science, cognitive science, neurology and psychiatry. It's progress is tracked in numerous dedicated journals, organised by a centralised scientific society (The Association for the Scientific Study of Consciousness) and its research outputs presented each year at various large and well attended conferences[16].

Here, I will briefly outline what I take to be the two central developments or broad research programmes in the contemporary neuroscience of consciousness which continues to comprise the majority of current research in the field: first, the construction of the *Neural Correlates of Consciousness* framework which serves as the starting point of the modern science of consciousness and second, the more recent attempts to develop a number of unified and comprehensive empirical *theories* of consciousness which purport to account for and explain the data emerging from the NCC framework in a principled and systematic manner[17]. More detailed discussions of this work, where applicable, are included later in the thesis. The purpose of this overview is to illustrate the extent and maturity of the research produced as part of the neuroscience of consciousness such that the claim serving as the starting point for this thesis is made salient in light of our previous discussion. This is the observation that, for the most part, contemporary literature on the metaphysical nature of consciousness fails to take into account (or even acknowledge) this large body of empirical research on the neural underpinnings of consciousness - a methodological thesis often supported by the further claim that this work is, in an important sense, *neutral or unilluminating* with respect to these metaphysical questions. This conspicuous claim about

[16]Recent overviews of the field as a whole and its progress can be found in Boly et al. (2013), Block et al. (2014) and Seth (2010). For the ASSC, see: http://theassc.org.

[17] Other prominent research which falls within this field which I omit from this overview includes work on specific contents of consciousness which relate to conscious selfhood (embodied, social and otherwise), volitional agency, emotion, attention and dreaming and related debates on the role of consciousness is social cognition and its distribution in infants and non-human animals (Boly et al. 2013).

the justification for the standard methodology adopted in the metaphysics of mind (and its implications) demands further examination.

2 1 The Neural Correlates of Consciousness

The neuroscience of consciousness, at least in its contemporary manifestation, starts from the construction and development of the *Neural Correlates of Consciousness* (NCC) research programme in the early 1990s. The NCC framework, introduced in a series of papers and books by Francis Crick and Christof Koch in response to the philosophical literature, and subsequently given a now-standard conceptual characterisation in Chalmers (2000), aims broadly toward identification of the *minimal neural conditions sufficient* for consciousness[18]. Since its introduction and formal characterisation by Chalmers, the NCC research programme has gone through numerous methodological, experimental and conceptual changes[19]. It is now generally agreed, however, that the NCC framework is constituted by two distinct approaches to studying consciousness, which can be distinguished in virtue of the aspect of conscious experience it investigates.

The first approach, referred to as content-specific, or 'building block' approaches, aims to identify the neural correlates of specific conscious *content.* Researchers working within this paradigm aim to identify and describe the neural substrate or mechanism that correlates with consciousness of particular intentional objects, such as faces, houses etc. and has been traditionally studied via report-based visual paradigms such as binocular rivalry, interocular suppression and various visual masking techniques[20]. Whilst these paradigms were integral to the early pursuit of the NCC framework , they have recently been superseded by modified paradigms which aim to "screen out" irrelevant neural activity relating to selective attention, self-monitoring or report, that precede or follow NCCs for specific contents [21]. Standing in opposition to this approach, the second line of research

[18] Historical precursors to the NCC framework can be found in work of british psychiatrist Henry Maudsley (1887) along with Herzen (1886) and Foster (1990) - see Michel (unpublished). For the initial development of the NCC programme see Crick & Koch (1990, 1995, 1998, 2003), Crick (1995, 1994) and Koch (2004). The various recent experimental and conceptual developments within the NCC programme over the last twenty years are nicely illustrated in Rees et al. (2002), Tononi & Koch (2008), Koch et al. (2016) and Howhy & Bayne (2015, 2016). The NCC programme as described here is crucial as it serves as the starting point for the argument in support of the neutrality assumption (see Chapter 2).

[19] See Noe and Thompson (2004a, 2004b, 2007, Searle (2005), Howhy (2009) for standard objections to the programme and the response in Metzinger (2000).

[20] See Tong et al. (1998), Logothetis et al. (2002), Tsuchiya and Koch (2005), Breitmeyer and Ogmen (2000). By way of explanation, in bincolular rivalry paradigms, distinct stimuli are presented to the eyes of a conscious subject which causes conscious experience to shift between the different stimuli every couple of seconds. Given that the stimuli remains constant in binocular rivalry paradigms, it is thought that these paradigms reveal, via fMRI, the content NCC for the specific content under study.

[21] See for example, Miller (2014), Aru et al. (2012) Tsuchiya et al. (2015). These so-called "no-report" and similar paradigms have led to a change in the purported location of content-specific NCCs from a fronto-parietal network (although to what extent this is implicated in content-specific NCCs remains contested) to posterior cortical areas (Koch et al. 2016).

within the contemporary neural correlates of consciousness framework aims to identify the neural correlates for a creature's overall *state* of consciousness (also called level, full or unified field approaches). While researchers working on content NCC(s) aim to identify the neural conditions minimally sufficient for a specific content of conscious experience (' a face' and so on), researchers looking to identify the state NCC(s) aim to identify the conditions minimally sufficient for having *any* conscious experience at all, irrespective of specific contents of a given conscious experience[22]. Defined formally by Howhy (2009:429) the NCC research programme can thus be characterised as a conjunction of the following approaches to studying consciousness, currently in progress:

1. *NCC for conscious content:* the minimally sufficient neural conditions for a specific (mostly representational) content being conscious rather than not being conscious.

2. *NCC for states of consciousness:* the minimally sufficient conditions for a creature's being in an overall conscious state rather than an overall unconscious state.

Crucial to the operational definition of the NCC programme just outlined is the notion of *minimally sufficient* conditions. Introduced initially by Chalmers (2000, see also 2010;chp3), this characterisation is designed to allow for the possibility of there being multiple NCCs for a given state or content (sufficiency), while capturing the idea that the search for NCCs must be the search for the central or core aspects of the system which are sufficient for consciousness - as opposed to the identification of a broader, more general system which is likewise sufficient (minimal)[23]. This is how the contemporary neural correlates of consciousness research programme has and continues to proceed.

2 2 Theories of Consciousness

Thirty years on from its introduction, work produced as part of the Neural Correlates of Consciousness research programme continues to dominate much of the scientific literature on consciousness. It is not exhaustive of the field, however. Prompted by the growing need to integrate the NCC data relating neural systems and conscious content and conscious level respectively into a systematic and unified account of

[22] The state-based approach is thus usually studied using experimental paradigms which exploit differences between cases where consciousness is present (as in healthy awake volunteers - HAWs) and where it is typically absent, typically, in the altered levels or disorders of consciousness which arise as result of brain damage (UWS patients) and seizures, or in patients under anesthesia or in dreamless sleep. Key studies include Laureys et. al (2004, 2005, 2014), Owen et al. (2006) Massimini et al. (2005), Brown et al. (2010) Siclari et al. (2014).

[23] C.f Fink (2016), who argues that the operational definition of the NCC ought to go beyond sufficiency and stipulate the *necessity* of the neural conditions or mechanisms at the level of phenomenal types.

consciousness and its cognitive function, recent work in the neuroscience of consciousness has been concerned with the production of numerous theoretical frameworks which aim to go beyond this early NCC research toward the production of comprehensive *theories* or models of consciousness (Boly et al. 2013). As is somewhat expected at this relatively early stage of research, a large number of (supposedly) competing theories and frameworks have been put forward. Out of these, four contemporary theories in particular have received sustained attention: Global Workspace Theories (Baars 1998, 2012 Dehane, Kerzberg & Changeux 2001), Higher Order Theories of consciousness (Rosenthal 2005), Tononi's Integrated Information Theory (Tononi 2004, 2008, 2012, Koch and Tononi 2008), and most recently, the attempt to extend and apply the popular prediction error minimisation framework (PEM) of brain function developed in cognitive science to consciousness (Hohwy 2012, 2013, Clark 2013, 2016)[24].

Whilst their proponents typically claim that these theories are comprehensive and exhaustive theoretical frameworks, it has been argued that as things currently stand, these are best understood with respect to the aspect(s) of consciousness - beyond the phenomenal - that each given theory takes as its central or starting explanatory target, given that these often diverge. For example, it has been claimed that Higher Order and Global Workspace theories, along with current research paradigms studying PEM approaches to consciousness are best characterised as targeting conscious *content* and notions of conscious *accessibility* (where this access claim applies particularly to Global Workspace Theories). This contrasts with the aims of Integrated Information Theory, whose central tenets are often motivated - along with the explicit desire to tackle the hard problem head on - by the clinical study and account of the difference between conscious *levels* and identification of the NCCs of state-consciousness[25]. The relationship between these theories and their explanatory goals however, especially with respect to the potential unificatory power of PEM frameworks, continues to be a matter of current empirical and conceptual debate[26].

---

[24]Other prominent accounts include virtual reality theories (Revonsuo 2000) Recurrent or 'reentrant' processing theories (Edelman 1989; Lamme 2006), the operational architectonics model (Fingelkurts 2009) along with many others.

[25]Recent IIT research is discussed in chapter 2, however discussions of PEM in this context can be found in Havlik et al. (2017), Bucci and Grasso (2017) and Hohwy (2012, 2013), Clark (2013, 2016). For Higher Order Thought theories and its variations (Gennaro 1996, Rosenthal 1997, 2005), (Carruthers 2000), Lycan (1996) and Lau and Rosenthal (2011) and empirical support for this view: Weiskrantz (1997), Dienes (2008) and Lau and Rosenthal (2011) (c.f. Block 2007, 2009). For GWS see originally Baars (1998, 2005) followed by Dehaene and Naccache (2001), Dehaene and Changeux (2004, 2005, Shanahan 2008).

[26] This characterisation of the relation between competing theoretical frameworks is given in Seth (2017) and Howhy (2016). Whether these are genuinely competing (that is, mutually exclusive) accounts of consciousness, or whether their central claims and experimental evidence can be subsumed under a broader framework remains an open question, hindered by the current lack of precision and available testable predictions these frameworks currently generate (Boly et al. 2013, Howhy 2016). For some specific proposals for combining PEM approaches and GWS and IIT however, see (Howhy 2013 pp 211-214) and (Bucci & Grasso 2017).

3 The Neutrality Assumption

As should be evident from the overviews just given, contemporary research in the neuroscience and metaphysics of consciousness both operate in widespread *methodological independence* from one another. That is to say, researchers working within these fields take themselves to be pursuing very different lines of work - albeit on the same subject - whose parallel developments have little, if any, immediate significance for their field. Within the neuroscience of consciousness, this can be seen right from the beginning, in the motivations given for the initial construction of the NCC programme. The objective in this case was clear, namely, to set aside the hard problem of consciousness dominating philosophical discussions of consciousness at the time, which asked *why and how* a given set of physical processes can give rise to consciousness, and instead work to identify the minimum set of neural processes with which it is systematically correlated[27]. This distinction between what we might call the 'why' and the 'what' questions of consciousness as a way of understanding the distinction between the metaphysical and neuroscientific research programmes remains influential in the contemporary scientific psyche[28].

My primary concern however is the philosophical dialectic, whose widespread methodological independence from neuroscientific work on consciousness is unmistakable. Whilst philosophical contributions have undoubtedly been integral to the construction of the science of consciousness, and continue to play an active role in its conceptual development, the substantial empirical work on consciousness which these philosophical contributions target have failed to influence, and are not reflected in, the metaphysical debate on consciousness which has developed alongside it. Both in the construction and subsequent defense of competing ontological theories - varieties of physicalism, non-physicalism and illusionism alike - the focus has been almost exclusively on discussion of various a priori arguments (the exclusion argument, anti-physicalist arguments and the phenomenal concept strategy, the combination problem - the list goes on). The metaphysical debate over consciousness' ontological status has stayed silent on these recent empirical developments. Furthermore, given the current pursuit of the naturalistic non-physicalist research programme along with recent a priori debates over the viability of illusionist approaches to consciousness, this methodological practice looks to continue.

---

[27] See Crick and Koch (2003;1): "It appears fruitless to approach this problem [the hard problem] head-on. Instead, we are attempting to find the neural correlate(s) of consciousness (NCC), in the hope that when we can explain the NCC in causal terms, this will make the problem of qualia clearer".

[28] See for example, Seth (2010): "Perhaps the key factor in the transition to scientific legitimacy was the realization that it may not be necessary to explain *why* consciousness exists in order to begin to unravel the physical and biological mechanisms that underlie its various properties. After all, physicists have laid bare many mysteries of the universe without accounting for the brute fact of its existence".

This observation is striking. The philosophers working on the metaphysics of consciousness are obviously well aware of this neuroscientific research. So what explains the absence of discussion of this empirical work in mainstream metaphysical debate? A natural thought is that it might be unreasonable to expect metaphysical theories of consciousness, as the product of a soley a priori enterprise which is concerned with a set of more abstract and fundamental questions regarding the nature and metaphysical constitution of consciousness, to be sensitive to these kinds of empirical considerations. If this were the case, then the lack of acknowledgement and attention to the recent neuroscientific work on behalf of metaphysicians of mind would make sense. However, a brief look at the historical developments pertaining to causal closure outlined above suggests that this is clearly not the case. With causal closure, we have a clear case in which empirical developments, namely, those implicated in the inductive arguments for the causal closure principle - one which, incidentally, is based on findings from *modern neuroscience* - have had a profound effect on the dialectical landscape in the metaphysics of mind. Here, a posteriori developments - both in the special sciences as well as physics - changed the theoretical set up completely, being responsible for the dominance of physicalism as the now default metaphysical view, the almost blanket rejection of interactionist dualism and, perhaps most interestingly, the subsequent demand on competing non-physicalist theories to accommodate the causal closure principle within their frameworks[29].

It is not the empirical, a posteriori nature of the neuroscientific research programme which stops it from entering mainstream contemporary metaphysical discussions. So what is it, then? What justifies the claim that we can, in simple terms, explicitly take on board everything neuroscience is telling us about the neural basis of consciousness yet end up with radically different metaphysical accounts of its nature? The justificatory explanation often given is as follows. That, beyond the developments pertaining to causal closure, the recent empirical work produced as part of the neuroscience of consciousness is, in an important sense, *neutral or unilluminating* with respect to these metaphysical issues. That is, it is claimed that the competing metaphysical accounts of consciousness are *equally compatible* with the work emerging from contemporary neuroscience - in a manner in which interactionist dualism was *not* with respect to the relevant work in the case for causal closure - given that this research is, metaphysically speaking, neutral or unbiased. On this view, the exclusively a priori nature of contemporary debate, despite the construction of the new science of consciousness, has a simple

[29]That is, those who reject physicalism do not tend to do so on the basis of rejecting the empirical case for causal closure - for example, by claiming that the distinctly a priori nature of metaphysical inquiry rids us of the demand to accept causal closure - but rather on the basis that the causal closure principle can instead be accommodated within alternative non-physicalist metaphysical frameworks (see for example Goff 2017). A more detailed account of the meta-metaphysics of mind which is suggested by this discussion is offered in Chapter 3.

explanation: that the work produced from this research field thus far is silent with respect to the metaphysical debates over phenomenal consciousness and its ontological status.

This neutrality claim is implicit in much of the contemporary philosophical dialectic above. As suggested, it is evidenced in the continued reliance on a priori considerations to settle the ontological disputes at the heart of current debate, along with the view that both modern non-physicalist theories and illusionist approaches are broadly consistent with a neuroscientific view of the mind and consciousness. However, the view can also be found explicitly stated in a number of recent papers. In the interests of clarity, and to make the target of this thesis clear, I quote three instances of this here[30]:

In her forthcoming discussion of the perceived ontological stalemate holding between physicalists and their opponents, Katalin Balog writes:

"It is unlikely that this stalemate can be broken by empirical evidence either. We have good reason to think that non-interactionist property dualism and physicalism *are equally compatible with all empirical evidence.* ... Of course, it could turn out that the physical is not causally closed and they are both wrong on this issue, but nothing so far points strongly in this direction"

"For physicalism to be true, phenomenal properties must have "neural correlates"....Because the non-interactionist property dualist – like the physicalist –believes that the physical is causally closed, *she is equally committed to the existence of such neural correlates."*

(Balog, in press;19; emphasis and [] added).

From within recent philosophical contributions to the science of consciousness, a similar claim can be found in Tim Bayne and Jakob Howhy's recent review and commentary of developments in the NCCs programme:

"The disagreement about the metaphysics of consciousness has little direct bearing on the NCC project, for all that requires is that certain neural states 'underlie' consciousness"

(Bayne and Howhy 2016;4).

[30] However see also Wu (forthcoming;7), along with the more comprehensive discussions in the papers introduced in Chapter 2.

And finally, in the research statement outlining the objectives of the New Directions In the Study of Mind, the the recent research project at the University of Cambridge which has funded the development of naturalistic non-physicalist frameworks, Tim Crane and colleagues write that:

"The current project rejects this assumption[that the empirical study of consciousness demands the truth of physicalism]. It will maintain that the scientific investigation of the mind—by psychology and neuroscience—does not require that physicalism is true. One of the distinctive features of this project, then, is the *combination of a skeptical attitude to physicalism with a fully scientific approach to the mind*".

(Crane et al. 2015; 3; emphasis added).

4 The Dialectical Function

The dialectical situation then, is as follows. Contemporary metaphysicians of mind - along with many neuroscientists working on consciousness - adhere to the following assumption:

The Neutrality Assumption: Work produced as part of the neuroscience of consciousness is neutral, or metaphysically unilluminating, with respect to theory choice and construction in the metaphysics of consciousness.

This explains the acceptance of, and justifies the following claim:

Robust Methodological Independence: The metaphysics of mind - concerned with construction and theory choice between competing metaphysical accounts of consciousness - ought to operate in widespread independence from empirical developments in the neuroscience of consciousness.

It is important to note that this latter methodological claim is *robust* -it applies to *all* of the relevant neuroscientific and metaphysical work and theories (beyond closure) and is thought, at least *prima facie,* to have some degree of temporal infragility[31]. If broadly correct, this suggests the following, that the neutrality

[31]That is, the view seems to be not only that *current* neuroscientific research on consciousness is neutral in this way but also, that this was so in the past and is also likely to remain the case in the near future - such that metaphysicians are justified in not continually looking for emerging empirical work which might bear on the relevant ontological questions (C.f. Chp2,3).

assumption is crucial for contemporary metaphysics of mind, and does significant philosophical work in current debate, in virtue of justifying its standard methodological practices viz. those which fail to engage in sustained consideration of emerging empirical work on consciousness and its neural basis, and take such an approach to be justified.

5 Overview of Subsequent Chapters

The mandate for the thesis then is to address the following questions: (i) what argument(s) support this neutrality claim? (ii) are they convincing? And (iii) if not, what does this mean for robust methodological independence? Chapter 2 addresses (i) and (ii). Chapter 3 discusses (iii), and proposes an alternative methodological approach. Chapter 4 examines the applications of this alternative methodological framework and the implications this might have, when put to use, for contemporary metaphysics of mind.

Chapter 2: The Neutrality Argument

The main line of argument in support of the neutrality assumption can be traced back to David Chalmers (2002, 2010). This takes as its starting point the Neural Correlates of Consciousness research programme. After providing a further exposition of this programme and Chalmers' argument, I examine the recent development and elaboration on this line of argument proposed by Uriah Kriegel (*forthcoming*). Together, these form what I call the 'neutrality argument' in favour of the neutrality assumption examined in this thesis. Drawing on recent work in philosophy (Neisser 2012, Vernazzani 2015, 2016) and neuroscience (Revonsuo 2000, Tononi et al. 2014), I present and examine two related objections to this argument. Together, these aim to demonstrate that the success of the neutrality argument depends, in various ways, on an outdated and empirically implausible view of contemporary neuroscientific research and, in particular, its explanatory aims and practices. I conclude that, in the absence of available alternative arguments for the neutrality, assumption, this claim and, by extension, the methodological thesis that the metaphysics and neuroscience of consciousness ought to operate independently, lack adequate justification.

Chapter 3: Towards a Neuroscience-First Metaphysics of Mind

In the next Chapter I build on this discussion and the related work it draws upon to motivate and present an alternative 'Neuroscience-First' methodology for the metaphysics of consciousness. This takes as its starting point the recognition that the question at the heart of the thesis - namely, that of the correct account of the relationship between the metaphysics and neuroscience of consciousness - is an instance of the broader *meta-metaphysical* issue of the relationship between metaphysics and science more generally, and as such, ought

to be illuminated by recent developments in this field. After detailing two such developments viz. (i) the recent emergence of work in the metaphysics of neuroscience, and (ii) the prominent methodological naturalist analysis of metaphysics in terms of *inference to the best explanation* (L.A.Paul 2012), I use these to construct a case in favour of a novel methodological approach to debates over consciousness' ontological status. This takes as central the claim that a Neuroscience-First approach to the metaphysics of consciousness - on which the metaphysical commitments of neuroscience and its explanatory practices serve as the *mutual starting constraints* on a metaphysics of consciousness - is not only a potential alternative to methodological independence, but is demanded by acceptance of, and reflection on, the minimal claim that candidate metaphysical theories of consciousness be *empirically equivalent*. I conclude by using this discussion to reinforce and restate my case against Chalmers' neutrality argument, and situate this approach in the context of the recent arguments for naturalised metaphysics.

Chapter 4: An Application: The Neural Mechanisms of Consciousness

In the final Chapter, I demonstrate how this novel methodological approach can be put to good use in the metaphysics of mind. Taking a mechanistic interpretation of the Neural Correlates of Consciousness programme (Revonsuo 2000, Neisser 2012, Vernazzani 2015) introduced in Chapter 1 as a natural starting point, I argue that attention to the metaphysical commitments of such a programme can be used to draw a negative conclusion with respect to contemporary theories in the metaphysics of mind. More specifically, I motivate the claim that the standard account of the metaphysical dependency relation holding between neural mechanisms and the phenomenon they explain (provided by recent philosophy and metaphysics of neuroscience) provides the resources to construct novel case against of type-identity theory (Polger 2006, Polger and Shapiro 2016). I conclude by situating these arguments in relation to the more familiar concerns pressed against these accounts of consciousness viz. Those relating to to multiple realizability and mental causation. As tools to secure the *empirical equivalence* of metaphysical accounts of consciousness (a result of the Neuroscience-First Approach in Chapter 3), I argue that such considerations ought to be taken as importantly distinct from, and dialectically *prior* to such arguments, which remain even if these earlier problems can be successfully overcome.

Chapter 2:
The Neutrality Argument

In Chapter 1 I explained and motivated the claim that the neutrality assumption does crucial work in the contemporary metaphysics of mind in virtue of justifying the latter's methodological independence from the neuroscience of consciousness. But what reasons are there for thinking that this assumption and, by extension, the standard methodological approach adopted in current debate, are in fact justified? As we have seen, in all major treatments of the metaphysics of consciousness found in recent literature, the neutrality assumption has received little, if any, attention. Where it is stated explicitly, as in the passages outlined in the previous chapter, the view is often asserted with little or no explanation; a claim which is striking given that historical developments suggest that metaphysical theories ought to be sensitive to these sorts of empirical developments.

In this chapter I shall demonstrate that these recent statements of the neutrality assumption can be traced back to a single line of argument which, if successful, provides an adequate justification of the neutrality assumption. This appears first in contemporary dialectic within the series of papers David Chalmers uses to outline his prescriptive account of a science of consciousness (1996, 2000, 2004). Whilst this argument has gone largely unscrutinised, the need for its further support and elaboration has been recognised by Uriah Kriegel who gives the argument a renewed treatment in a forthcoming paper. This 'neutrality argument' takes as its starting point the *neural correlates of consciousness* (NCC) research programme which, as discussed in the previous chapter, has gone on to dominate much of the neuroscientific work on consciousness in the past two decades. After setting this argument out, I claim that this faces two serious objections, on the basis of which, I argue, we can infer its inadequacy, and conclude that - as matters currently stand - the neutrality assumption and, by extension, robust methodological independence, lack adequate justification.

1. The Neutrality Argument

1 1 David Chalmers' Neutrality Argument

The need for a neutrality argument, as opposed to merely asserting the neutrality *assumption*, was first recognised in contemporary dialectic by David Chalmers in the series of papers he used to outline his prescriptive account of a science of consciousness (1996, 2000, 2004). For Chalmers, the ultimate aim of a science of consciousness is to infer *fundamental principles* connecting third person neural and behavioural data about the brain and first person data concerning consciousness collected indirectly via verbal reports[32]. That is, despite making a strong case for taking the neural correlates of consciousness as its' immediate centerpiece, Chalmers does not think that the successful identification of the NCC(s) is the end goal of the science of consciousness. In his prescriptive account of the science of consciousness (2004, 2010) he proposes two further projects for a science of consciousness beyond the completion of the NCC programme. First, Chalmers proposes that the next step for neuroscientists after the identification of the NCC(s) will be to work towards *systematising* the connection between first and third person data suggested by the NCC programme. As described in the previous chapter, the NCC paradigm is usually conducted in a disunified and fragmented way, divided between two approaches and working within distinct experimental paradigms. Systemation as prescribed by Chalmers would thus involve the unification of these approaches and the NCC(s) they identify, such that, if successful, we would be able to test or predict aspects of conscious experience based on an examination of an organism's neurophysiology. The final step for a science of consciousness according to Chalmers, is to infer a number of *fundamental principles* from these systematic connections. While it is difficult to tell exactly what Chalmers' means by this, he argues that these principles will be simple, unified and, most importantly, maximally general in scope such that they apply to *all* aspects of *any* conscious system. For Chalmers, the production of these simple law-like principles is the ultimate goal for a science of consciousness.

What would the construction of these principles tell us about the metaphysics of consciousness? Chalmers argues that the identification of these law-like principles would not, by themselves, give us a completed theory of consciousness because these principles *leave open* or *underdetermine* a metaphysical account of consciousness. Viewed as correlations, these principles are compatible with numerous metaphysical theories of consciousness and thus, are neutral with respect to them. It is open for the physicalist, for example, to argue that the existence of these fundamental, systematic principles are indicative of identity relations obtaining between neural and conscious states. Yet the existence of the fundamental systematic principles connecting third and first person relata can equally be thought to support other metaphysical accounts of the relationship between consciousness and brain states; the principles could - as Chalmers himself suggests - be used to support the

[32] Cf. Dennett (2001) '*The Fantasy of First Person Science*'.

existence of psychophysical bridging laws connecting physical and nonphysical domains, or alternatively, the claim that physical processes *cause* consciousness, and so on.

In sum, the idea that Chalmers puts forward is that the fundamental principles - inferred eventually from the work compiled as part of the NCC programme - *leave open* or underdetermine a metaphysical account of consciousness. Taken as correlations, one can claim that neuroscience, in so far as it aims at these principles, is neutral with respect to metaphysics of consciousness:

"*For many purposes, the science of consciousness can remain neutral with respect to these philosophical questions. One can simply regard them as principles of correlation, while staying neutral on their underlying causal and ontological status*" (2010:47).

There are three important features of Chalmers' formulation of the neutrality argument. Foremost, is his account of the neutrality (of neuroscience) being a result of the neutrality of *correlational facts,* which are compatible with numerous underlying causal or ontological relations. Second, is that Chalmers' neutrality argument is intended to be applied to a completed science of consciousness, and is therefore based around principles which have not yet been uncovered (if, indeed, they can be). As such, it does not obviously bear on the question of whether *current* neuroscience is similarly neutral[33]. Finally, a striking feature of Chalmers' argument comes from its' fleeting position in his overall account of a science of consciousness, and the rigorous treatment of consciousness presented in his book. The argument is only introduced briefly at the end of the chapter outlining the projects and problems for a science of consciousness. Whilst it is suggested that the topic will be taken up in later chapters (2010, 47) after an extensive treatment of the NCC programme, the discussion moves straight on to the ontological debate on consciousness. The relationship between the neuroscience of consciousness and the metaphysics, and further defense of the neutrality view that is proposed, is left undiscussed. Given how far the neuroscience of consciousness as a field has expanded, and the crucial role this assumption plays in contemporary metaphysics of mind, the need for a justificatory argument to establish the neutrality assumption introduced by Chalmers is imperative.

1 2 Uriah Kriegel's Neutrality Argument

[33] However, as I will suggest below, we can construct a formulation Chalmers' neutrality argument which does so based on the generalisation of the above to all [current and future] neuroscientific work.

The task of providing a more comprehensive and up to date defense of the neutrality assumption has been taken up by Uriah Kriegel, who gives the argument its most extensive treatment in a forthcoming paper[34]. Observing similarly that the centerpiece of the science of consciousness is the NCC research programme, Kriegel's argument starts from the claim that the NCC research programme as outlined above is problematic from a philosophy of science perspective. The motivation for Kriegel's formulation of the neutrality argument is thus the thought that science places intellectual demands on neuroscience, such that a scientific study of consciousness framed only in terms of correlations - whose systematic existence is left unexplained - is unsatisfactory. That is, Kriegel claims, if the NCC programme is to be a properly scientific enterprise it must go beyond the *identification of the neural correlates* of consciousness toward the provision of *explanations* of their existence. To limit the scientific study of consciousness as the search for correlations would be to render the existence of the correlations brute and inexplicable which, Kriegel claims, runs contrary to scientific practice[35].

As such, Kriegel identifies six possible explanations for the correlation(s) identified by the NCC research programme, which he cashes in terms of constitutive or causal relations. An explanation for the correlation between a given conscious experience and its' neural correlate, neutral between state and content NCC approaches, will either be *causal* (consciousness is constituted by the NCC), r*everse causal* (the NCC is constituted by consciousness), *third causal* (consciousness and the NCC are both caused by some third element), *constitution* (consciousness is constituted by the NCC), r*everse constitution* (the NCC is constituted by consciousness) or *third constitutor* (consciousness and the NCC are both constituted by some third element). After laying these out as the most plausible explanations for the correlational facts identified by the NCC programme, Kriegel argues that each explanation maps onto a familiar position in the metaphysics of mind. These are as follows[36]:

Causation -- Naturalistic Dualism (*epiphenomenalism*)

---

[34] 'Beyond the Neural Correlates of Consciousness'. Forthcoming in U.Kriegel (eds.) *Oxford Handbook of the Philosophy of Consciousness*.

[35] As such, Kriegel seems to disagree with the qualification made by Chalmers that such correlations (when systematically reduced to fundamental principles) might reasonably be taken as *fundamental* or primitive, in a way analogous to other fundamental laws in physics. For if this was correct, then no further explanation of the NCC correlations would be required.

[36] I think there are a number of potentially serious problems with Kriegel's mapping of metaphysical theories with causal and constitutive relations in this way. To take one example, it is highly controversial to state that the metaphysical dependency relations used to characterise naturalistic dualism (nomological supervenience etc.), for example, are instances of diachronic *causal* relations (cf. Koslicki 2016, Bernstein 2016). Neither are we beholden to accept the claim that constitution relations entail *identity* relations (as seen in the familiar puzzles of material constitution e.g. in Lewis 1976). Here however, I grant Kriegel characterisation and focus on the implications of the argument, if successful, for the construction of a viable neutrality argument.

Reverse Causation -- Non-naturalistic dualism (*interactionist dualism*)

Third Causation --Neutral Dualism (*not defended*)

Constitution -- Physicalism (*a priori & a posteriori physicalism*)

Reverse Constitution -- Idealism (*reductive and eliminative idealism*)

Third Constitutor -- Neutral Monism (*Russellian monism*)

Kriegel's defense of the claim that the neuroscience of consciousness is neutral with respect to the metaphysics of consciousness follows from his assertion that, once these possible explanations are laid out, there is no *scientific* way of distinguishing or choosing between them. Broadly speaking, when assessing the viability of different explanatory hypotheses for a given phenomenon -viz. theory choice - science proceeds by first asking which explanatory hypotheses best accommodates the empirical data. In other words, the hypotheses are assessed for empirical adequacy. This is typically done by constructing experimental paradigms designed to exploit *discordant predictions* or observational consequences of the hypotheses under consideration. This allows researchers to distinguish between the competing explanatory hypotheses for a given phenomena, keeping those which best fit the newly acquired data[37].

Kriegel's neutrality argument then is that, in the case of the competing metaphysical explanatory hypotheses outlined above, we are unable to assess or choose between the explanatory hypotheses for the NCC in this way. Examining two 'empirical symptoms' or discordant predictions that might be used to distinguish between naturalistic dualism and physicalism - time lag and mediating mechanism respectively - Kriegel argues that there are both serious technological and philosophical barriers to testing these empirically[38]. In the absence of the empirical data which can be used to distinguish between dualism and physicalism, he concludes that there is no hope for a scientific resolution to the choice between the metaphysical explanatory hypotheses; the science of consciousness, constituted on this view by NCC research, is neutral with respect to the metaphysics of mind[39].

---

[37] See Chp3.3 for further discussion.

[38] This follows from the claim that diachronic causes usually *precede* their effects in addition to the claim that causes are usually mediated by *mechanisms* involving fine grained causal transactions, whereas constitutive relations typically do not. Considering time lag, Kriegel provides three problems (a) a technological barrier concerning temporal resolution required to test the hypotheses (b) the supposed unobservability of causation (c) the inability to control for measuring time. For causal mediating mechanism Kriegel argues that this is problematic in this instance as this does not apply to causal relations occurring at the fundamental level, which naturalistic dualism is concerned with. This thus cannot be used to distinguish empirically between epiphenomenalism and physicalism understood by Kriegel as involving causal and constitutive explanatory hypotheses. (Here, Kriegel seems to be implicitly relying on a mechanical account of causality Glennan 1996).

[39] Kriegel also includes a discussion of *theoretical* neutrality (22,24), but I leave this out of my discussion here.

Key to Kriegel's neutrality argument is his expansion of Chalmers account of neutrality as a matter of 'compatibility' of correlational facts with numerous metaphysical interpretations, to an account of the neutrality of these facts being as a result of their *empirical equivalence*. Kriegel's account of neutrality in this context can be summed up as follows:

Neutrality as empirical equivalency: A set of scientific facts can be said to be neutral with respect to a set of metaphysical theories (T_1, T_2, T_3.. T_x) iff that set of facts fails to produce *empirical evidence* (viz. a set of discordant predictions) on the basis of which we could positively distinguish between them.

Like Chalmers' argument, Kriegel's neutrality argument has a number of important features. First, it demonstrates that the neutrality exhibited between the neuroscience and metaphysics of consciousness is contingent and temporally fragile. According to Kriegel, it *just so happens* that there are obstacles to the production of relevant experimental paradigms which exploit the discordant predictions of the competing metaphysical explanations proposed for the correlation (which would allow us to choose between competing metaphysical theories). That is, as it stands, Kriegel's argument fails to establish a neutrality claim which is *robust*; the neutrality of neuroscience might change if more work was put into designing such paradigms, thinking up new discordant predictions that the metaphysical explanatory hypotheses might generate, or if and when neuroimaging and other experimental techniques improve in the future[40]. Second, another important feature of Kriegel's argument is his characterisation of scientific explanation exhaustively in terms of single *causal and constitutive* relations. While this is useful for Kriegel's purposes in so far as it limits the scope of discussion to a manageable number of hypotheses which can be later mapped onto metaphysical accounts, I think there are strong reasons to think that bare stipulation of constitutive and causal relations are neither exhaustive exemplars of (neuro)scientific explanation nor reflective of the notion of explanation at work in neuroscientific practice[41].

1 3 A General Outline

While there are a number of important differences in the formulations of the neutrality argument presented by Chalmers and Kriegel, they both exhibit the same general structure:

[40]Kriegel does admittedly express doubt at the prospect of this (22,25), but it is not altogether clear that this pessimism is justified. It seems difficult, for example, to predict in advance what the relevant technological developments will be, and thus, why we should rule out the testability of these competing metaphysical hypotheses in advance.
[41] See sections 2 and 3.

1. Neuroscience as it pertains to the study of consciousness discovers correlational facts.
2. Correlational facts are metaphysically neutral with respect to competing metaphysical theories of mind.
3. Therefore the neuroscience of consciousness is neutral with respect to the metaphysics of mind.

As described, both formulations of the neutrality argument start with the NCC research programme, characterised by its search for the identification of the neural correlates of conscious experience (that is, a set of correlational facts) . It is then the neutrality of *these* correlational facts which is used to establish the neutrality of neuroscientific practice with respect to the metaphysics of mind. This is supported either by an appeal to the broad compatibility of such facts with the divergent metaphysical accounts of consciousness (Chalmers) and/ or with the latter's empirical equivalence due to a failure to test for discordant predictions (Kriegel). Taken as the cornerstone for the scientific study on consciousness, this neutrality conclusion as it relates to the NCC programme is then generalised to include *all* neuroscientific work on consciousness, and thus the robust neutrality assumption outlined in the introduction is established. This general structure provides a firm starting point from which the viability of the neutrality argument in favour of the neutrality assumption, as it previously been presented, can be assessed.

2. Two Problems for the Neutrality Argument

There are strong *prima facie* reasons for accepting the second premise of the neutrality argument. The metaphysical neutrality of correlational facts is relatively uncontroversial, and Kriegel's forthcoming account of neutrality as empirical equivalence due to discordant predictions serves as a plausible example of how the premise can be defended under closer examination[42]. However, I shall argue that this is not true of the first premise of the neutrality argument as defended by Kriegel and Chalmers. In the remainder of the chapter I raise and discuss two objections which suggest that the claim serving as the first premise of the argument, namely,

[42] That being said, I also think there is a case to be made for a further examination of the second premise of the argument. If the neutrality argument is to be supported in the manner required to establish the neutrality assumption as it operates in contemporary debate, Kriegel's argument would require further elaboration and treatment. Minimally, the argument would require that it was established that *all* of the metaphysical explanatory hypotheses Kriegel outlines (assuming, not uncontroversially that these correspond to the metaphysical account of consciousness that is required, c.f. Footnote 5) are empirically equivalent in the way that Kriegel argues is true of physicalism and epiphenomenalism (that is, as it stands, Kriegel's argument is currently invalid given that he does not defend the empirical equivalence of all competing accounts). This seems to be a potentially fruitful line of research that is worth pursuing. However, given that my aim here is to discuss what I see as stronger objections to the neutrality argument, I do not pursue this objection to Kriegel's defense of premise 2 of the neutrality argument, which would add weight to my conclusion, any further here.

that neuroscience as it pertains to the study of consciousness discovers *correlational facts,* is false. These are as follows. First, that we have good reason to think that the correlational account of the neural correlates of consciousness programme which this argument relies on is mistaken. That is, an alternative and more plausible characterisation or interpretation of the programme exists according to which the NCC is best understood as the search for *causal or constitutive* facts which are not neutral in the manner required to establish the argument's conclusion. Second, I will argue that even this first objection can be resisted, the neutrality argument detailed by Chalmers and Kriegel fails because - as alluded to in the introduction - the neural correlates of consciousness research programme is no longer *exhaustive* of neuroscientific work on consciousness. If prominent empirical work exists which falls outside of this programme and, moreover, which similarly resists correlational analysis, then premise one is false. I detail a case study of work which meets these two conditions. I finish with a discussion of the large problem I take these two preliminary objections to be indicative of viz. a failure in previous discussions of this issue to appreciate the explanatory and metaphysical implications of neuroscientific work on consciousness, the denial of which, I shall later argue, is becoming increasingly implausible.

2 1 Are the NCCs Correctly Characterised as Correlational?

Premise one of the neutrality argument as I have outlined it relies on the claim that the NCC programme is correctly characterised as searching for *correlational facts* relating first and third person data. The first objection to be pressed to this neutrality argument then, is that there is strong case available against this original correlational characterisation of the NCC which, if correct, renders the neutrality argument unsound (Neisser 2012, Vernazzani 2015, Revonsuo 2000). This line of argument takes as its starting point the observation that current philosophical discussions of the science of consciousness and its relation to the standard metaphysical debates suffer from a lack of attention to philosophy of science and, in particular, to the influential account of *mechanistic explanation* detailed by contemporary philosophers of neuroscience.

An account of how cognitive neuroscience proceeds, and of its aims, norms and explanatory practices has been a key subject of debate in recent philosophy of science. In the last twenty years, philosophical accounts of scientific explanation have taken a "mechanistic turn" (Kästner 2017). Replacing previous deductive-nomological (Hempel & Oppenheim 1948) and unificationist accounts (Kitcher 1989) which proved to be unequipped to deal with the explanatory practices emerging in the special sciences, the now received

account of neuroscientific explanation is that proposed by the *New Mechansits*[43]. According to this view, neuroscientific explanation works by uniformly accounting for higher level cognitive phenomena, diverse behaviours such as memory, language and action, via the identification and detailed description of the *neural mechanism* which brings it about and, in doing so, situating such cognitive phenomena within the *causal structure* of the world. The standard new mechanist form of explanation is illustrated in the following diagram:

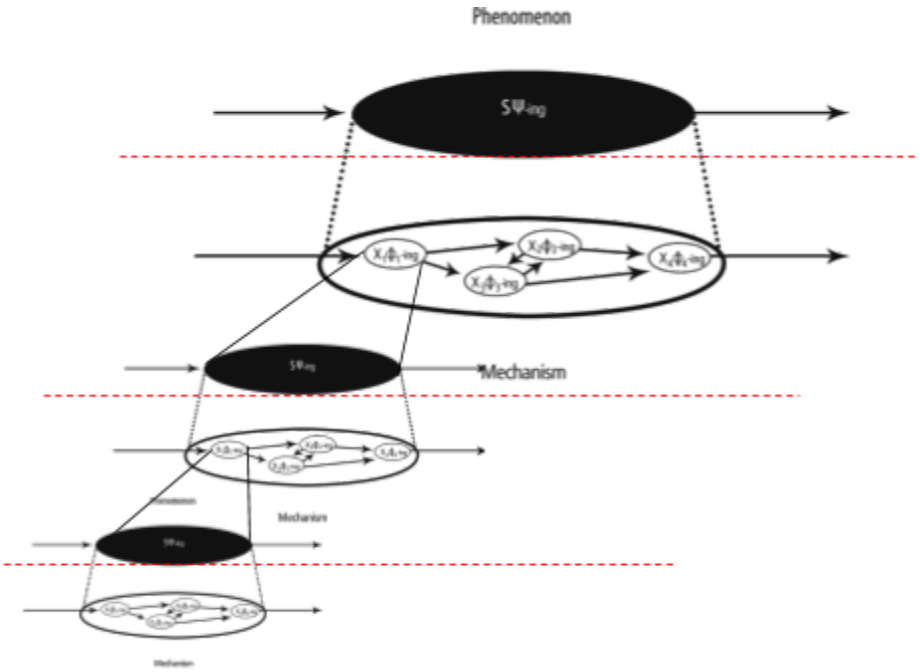


Figure 1 Mechanistic Explanation (adapted from Craver 2007;7)

In figure 1 the cognitive phenomena to be explained is represented by 'S's ψ-ing', where S refers to the mechanism as a whole, and ψ the property or behaviour to be explained[44]. According to the mechanistic account of explanation, this is explained by the temporal and spatial organisation of the activities of component entities (Xs' ϕ-ings) which make up the mechanism (Craver 2007;5-7). In the diagram, the X's represent the component entities of the mechanism, the ϕ's the activities which they perform and the arrows denote the causal activity between the acting component entities in the mechanism. In mechanistic explanation, the relationship between the explanandum (the higher level cognitive phenomenon) and the explanans (the neural mechanism) is said to

[43] New Mechanists include Machamer Darden & Craver (2000) Bechtel (2005, 2007), Darden (2008), Glennan (2005, 2017) Machamer (2004), Woodward (2002), Ilari & Williamson (2012). I concern myself here with the most influential account of mechanistic explanation put forward by Carl Craver (2005, 2007, 2013).
[44] Recent debate concerns the ontological nature of constitutive mechanistic phenomena. See Kaiser & Krickel (2016) for an overview and convincing argument for the claim that these are best understood as *object-involving occurents*.

be one of *constitution* (indicated by the black dotted lines in figure 1). That is, the acting entities which constitute the mechanism (Xs' ϕ-ings) are spatially and temporally *contained within* the phenomenon (S's ψ-ing) . Indicated by the way in which phenomena are often said to be explained by their 'underlying mechanism' which 'exhibit' or are responsible for them, this form of explanation in cognitive neuroscience is known as *constitutive mechanistic explanation*.

According to this account of explanation, cognitive mechanistic phenomena are constitutively explained by their being situated within a hierarchy of *levels* or 'nested mechanisms', defined locally in terms of the component acting entities of a given mechanism (represented by red lines in figure 1). That is, once the component acting entities of the mechanism for a given phenomenon have been identified, these can then in turn be explained via the identification and description of *their* underlying mechanism, by the same method[45]. Craver puts forward the following criterion for the individuation of mechanistic levels: ''X's ' ϕ-ing is at a lower mechanistic level than S's ψ-ing if and only if X's ϕ-ing is a *component* of the mechanism for S's ψ-ing'' (189). How is it that components (Xs' ϕ-ings) of a given mechanism for S are determined? Craver outlines two conditions for the constitutive relevance of component entities of mechanisms(153-157): (i) spatiotemporal parthood; X must be a spatiotemporal part of the system whose behaviour is to be explained (Krickel 2017). And (ii) Xs' ϕ-ings and S's ψ-ing must be *mutually manipulable* viz. A part is constitutively relevant component in a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the behavior of the component *and* one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole[46].

A number of philosophers have recently argued this prevalent form of mechanistic explanation just outlined places interpretational demands on a characterisation of the NCC programme, demanding its recharacterisation. In his paper 'The Neural Correlates of Consciousness Reconsidered' (2012) Joseph Neisser argues that any discussion of the NCC programme ought to reflect this prevalent account of neuroscientific explanation[47]. Motivated again by the idea that contemporary discussions of the NCC programme suffers from a lack of input from recent philosophy of science, Neisser's' objection stems from what he sees an implicit

[45] The standard example of this is that of spatial memory, illustrated in a rat's navigation of the Morris Water Maze (Krickel 2017;3, Craver 2007; 165-70).

[46] The viability of the *mutual manipulability criterion*, characterised in terms of Woodwardian interventions (153), has been the topic of recent debate (Leuridan 2012, Baumgartner & Gebharter 2016, Romero 2015). The worry here is that there is a tension or inconsistency in Craver's account insofar as interventionism is an account of *causation*. See Romero (2015) Baumgartner & Gebharter (2016) and Krickel (*forthcoming*) for ways of resolving this inconsistency and retaining the constitution relation along mutual manipulability lines. It is this ambiguity between constitutive and causal relations which ultimately leads Neisser (below) to his causal characterisation of the NCC. I return to this in chapter 4.

[47] See also Revonsuo (2000), Vernazzani (2015), Seth (2009). For ease of exposition, I focus of Neisser's argument here.

commitment in Chalmers' writing to an outdated form of scientific explanation viz. an account which takes the crucial explanatory aim of neuroscience to be the discovery of laws from which deductive arguments can be constructed[48]. According to Neisser, once the mechanistic explanatory aims and practice of neuroscience are taken into account, the NCC programme ought to be understood as working towards an *explanation of consciousness* (which applies to both content and state approaches) via the description and identification of the *multi-levelled neural mechanisms* responsible for its production.

That is, whilst the correlational interpretation looks to be correct on Chalmers' description of the the science of consciousness - on which neuroscience ought to aim at formulation of a number of fundamental law-like principles - it does not make sense on, or follow from, a *mechanistic account* of neuroscientific explanation. On what Neisser calls a 'causal mechanical' interpretation of the NCC, researchers are not identifying a set of correlational facts which connect first and third person data - that is, correlations between neural activations and conscious content - but instead are, as above, best understood as localising the working parts of an underlying multi-level *causal mechanism* which exhibits consciousness. Neisser concludes with the following revised definition of a content NCC:

"An NCC can be defined as a minimal neural system N such that states of N are underlying **causes** of a measurable change in consciousness, where a given state of N, as the causally relevant component of an embodied mechanism, is a mutually manipulable INUS condition for the specified aspect of the conscious state" (2012, 689).

Neisser supports this conclusion via a case study examination of binocular rivalry paradigms commonly found in the content NCC approach. According to Neisser, attention to the details of this experimental paradigm supports his causal characterisation insofar as the neuroscientists working within this paradigm are identifying the neural activations which *make the difference*, in the Woodwardian causal interventionist sense utilised by the new mechanists, to the experience of the preferred stimulus under examination in the study (683,686). That is, Neisser claims, not only is the mechanistic interpretation of the NCC *theoretically plausible* given the

[48] While he does not discuss this explicitly, what Neisser seems to have in mind - and what seems plausible given Chalmers prescriptive account of the science of consciousness above - is that Chalmers' is committed to the traditional Covering-Law model of explanation (Hempel, 1965) according to which scientific explanations are arguments from premises describing the laws of nature (Chalmers' fundamental principles) and antecedent conditions to a conclusion describing the explanandum. The covering-law model faces a number of standard objections, and is now widely believed to be inadequate (See Craver (2007, chp2 S.4 for a summary of these objections).

prevalence of mechanistic explanation in neuroscience, it is also supported by the actual experimental practices of researchers searching for the neural correlates of consciousness[49].

According to Neisser then, premise one of the neutrality argument is false as neuroscience, qua the neural correlates of consciousness research programme, is best understood as searching for a set of diachronic *causal* facts relating third and first person relata. If correct, this has serious ramifications for the proponent of the neutrality argument. Not only does it render the current argument unsound, but it also casts serious doubt on the prospects for constructing an argument in support of the neutrality assumption which takes this causal characterisation as a starting point. Causal relations, unlike correlational relations, come with metaphysical implications and commitments. To take one example, causal relations are generally thought to be *irreflexive* viz. requiring the existence of distinct causal relata[50]. If Neisser is correct then that mechanistic explanation demands a causal characterisation of the NCC, this would - following standard accounts of such positions - appear to rule out a variety of physicalist (constitutive or identity) accounts of the relation between the NCC and conscious experience and conversely, support prima facie a sort of *dualist position* (if Kriegel's causal characterisation of dualism is correct).

Whilst I agree with the broad motivation for Neisser's argument that the correct characterisation of the NCC programme ought to plausibly reflect the dominant forms of explanation in neuroscience (assuming that experimental practices support and don't point against this), I am not so sure that this demands a causal - as opposed to constitutive - interpretation. I return to this issue in the final chapter. For the purposes of my argument here however, it is enough that a mechanistic interpretation of the NCC - irrespective of whether it is best interpreted in terms of diachronic causal or synchronic constitutive facts - provides us with a characterisation of the facts uncovered by the NCC programme which are (a) *non-correlational* (thus falsifying the first premise of the neutrality argument) and perhaps more importantly (b) *not metaphysically innocent.* While only (a) is needed to refute the argument detailed in section 2, (b) is crucial; suggesting that the prospects of constructing an argument to justify the methodological independence exhibited by contemporary debate is slim. Can this line of argument be resisted? In the third and fourth chapters, I offer a more comprehensive case in favour of this claim. For now, however, I'll raise another objection against premise one of the neutrality

[49] Talk and discussion of the non-correlational neural mechanisms 'underlying' 'responsible' or 'generating' consciousness within the NCCs programme and beyond which support this claim are rife. See, for example, Anil Seth's 'Explanatory Correlates of Consciousness' (2009). In chapter 4 I return to, and give a more detailed case in support of this mechanistic interpretation of the NCC, and its experimental support, in the context of the more recent 'no-report based' experimental paradigms used in NCC work and constitutive mechanistic explanation.
[50] Psillos (2008).

argument which aims to demonstrate that it is *incomplete*. This applies even if the mechanistic interpretation of the NCC programme just outlined can be resisted.

2 2 Is the NCC Exhaustive of Neuroscientific Work on Consciousness?

In order for the neutrality argument to succeed in establishing a neutrality assumption which is *robust* - that is, applying to *all* neuroscientific work on consciousness - and thus justifying the robust methodological independence outlined in the introduction, it needs to be the case that the neural correlates of consciousness research programme is not only paradigmatic of, but *exhausts* contemporary empirical work on consciousness. That is, if there is prominent work on phenomenal consciousness which (i) falls outside of the NCC programme and (ii) can not be characterised in terms of the search for metaphysically innocent correlational facts, then premise 1 of the neutrality argument provided by Chalmers and Kriegel is false. Here, I will argue that there is at least one influential body of work in contemporary neuroscience which meets these two conditions. This is Tononi's Integrated Information Theory (IIT) (Oizumi et al. 2014, Tononi & Koch 2015, Tononi et. al 2016). After providing a brief outline of the theory, I'll identify what I take to be its central - non-correlational - ontological commitments.

One the leading and most influential neuroscientific theories of consciousness, IIT is motivated by the explicit aim to 'address the hard problem of consciousness in a new way' (Tononi et al. 2016;450). It purports to do so via adoption of a novel top-down or 'axiomatic approach'. Starting with the identification of a number of 'self-evident' claims about the essential nature of conscious experience - the so-called *phenomenological axioms* - proponents of IIT attempt to construct a neuroscientific theory of the physical basis of conscious experience by deriving from these self evident claims, a number of *physical postulates* which explain how these aspects of our conscious experience could be realised through a physical system like the brain. The theory has gone through several modifications[51], however in its most recent formulation (references above) IIT makes use of five such phenomenological axioms: (i) intrinsic existence (consciousness exists from its own 'intrinsic perspective' independent of external observers) (ii) composition (experience is structured in the familiar sense of allowing for various phenomenal distinctions - spatial, visual and otherwise) (iii) information (conscious experience is *specific*) (iv) integration (experience is unitary) and (v) exclusion (consciousness has determinate contents)[52]. The physical postulates - which again, attempt to explain how these axioms are brought about in the brain - are numerous, however, each is thought to follow *a priori* via deductive inference from the self evident

[51] Tononi (2004), Tononi (2008), Ozumi et al. (2014).
[52] C.f. Cerullo (2015).

phenomenological claims, a methodology which purportedly gives IIT its epistemic credentials[53]. These five axioms and their physical postulates lead proponents of IIT to their central claim: that conscious experience *is maximum integrated information* in a system (**Φ**). That is, according to proponents of IIT that the quantity of consciousness present in a system is identical to the amount of information generated by a complex of elements above and beyond the information generated by its parts[54].

In his recent critique of IIT, Tim Bayne (forthcoming) sets out the theory's explanatory aims and ambitions, which he takes to be threefold (i) being a theory of subjective experience (ii) being a *reductive theory* with ontological and epistemic commitments and (iii) being comprehensive, applying to each and every instance of conscious experience[55]. For my purposes in assessing the first premise of the neutrality argument, the first two commitments of IIT are crucial. First, IIT is explicitly a theory of phenomenal consciousness. This is important. While there may be other prominent neuroscientific work on consciousness which resists correlational analysis - such as Dehane's Global Workspace Theory - this frequently falls foul of the objection that it fails to account for or explain the phenomenal, as opposed to mere access, aspects of consciousness Block (2009).In contrast, IIT amounts to prominent neuroscientific work on phenomenal consciousness, which also falls outside of the NCC programme, and in doing so, meets the first condition outlined above. Second, IIT is an account of the fundamental nature of consciousness and comes with specific metaphysical commitments. That is its proponents are not claiming that conscious experience is merely *correlated* with maximum **Φ** but rather, the more significant claim that consciousness is to *be identified* with a given quantity of **Φ** in a system. While the specific metaphysical commitments of IIT - either physicalist or not - are currently in dispute, all that is required

[53] To take (ii) composition as one example (see also Mindt 2017; Tononi and Koch 2015). Here, the purported self-evident claim that experience is essentially compositionally structured (within my visual field I can for e.g. currently distinguish between my[laptop] sitting on the [desk] [in front] of me, the [red] [coffee cup] to the [side] etc. - creating a composition of phenomenal distinctions) demands that the physical system instantiating this aspect of phenomenology must itself be structured, where this is defined as 'subsets of elements (composed in various combinations) must have cause-effect power on the system' (Tononi & Koch 2015;7) An informative summary of the physical postulates of IIT can be found in Tononi et al (2016; 450-452). The move from axioms to postulates in each case can and has, along with the self evident nature of these phenomenological claims been questioned (see Switzgabel 2015; Bayne forthcoming;). Given that my aim here is draw out the ontological commitments of IIT, *if successful*, I don't discuss these here.

[54]Or more precisely, that every given conscious experience is identical to a *conceptual structure* - the set of cause-effect repertoires specified by a neural mechanism with maximum **Φ** (2016; 452). Defined as such, this identity claim provides an explanation of conscious content - which is said to correspond to the *form* of this structure - in addition to conscious level (which corresponds to its maximum **Φ**). The credibility of IIT as a theory of consciousness has come as a result of its explanatory and predictive power (c.f. Pautz *in draft*). The theory's central claim has since been tested, for example, at the level of individual subjects - ranging from healthy awake volunteers to brain damaged patients, using the Perturbational Complexity Index (PCI), which acts as a proxy for an empirical measure of integrated information in a system (2016;459, 2018;chp5). However development of further empirical methods for testing IIT's predictions is currently underway (Boly et al. 2015, and the *forthcoming* Entropy: Special Issue 2018).

[55] "for if- as its advocates claim - consciousness *just is* integrated information, then any system with integrated information must be conscious and any conscious system must exhibit integrated information" (2).

here is the following: that IIT amounts to an *ontological,* as opposed to correlational, neuroscientific thesis about the nature of phenomenal consciousness. In sum, insofar as IIT meets both conditions introduced above, the first premise of the neutrality argument is false: neuroscience as it pertains to the study of consciousness is not currently limited to the discovery of correlational facts as stipulated by the proponents of the neutrality argument. The neutrality argument is unsound.

Does this claim reflect ontological commitments internal to the theory itself, or just the somewhat naive metaphysical claims of its leading proponents? If this latter analysis is correct, an objection to the claim just made might be posed as follows viz. That we can grant the claim that IIT's *proponents* take their theory to have certain non-correlational ontological implications, but argue that, metaphysically speaking, it is open to the metaphysician of mind to view IIT's central claim relating phenomenal consciousness to **Φ** as *leaving open*, in much the same way as the correlational facts produced by the NCC programme, which precise dependency relation - identity, realisation, nomological or metaphysical supervenience etc. - the central relational claim of IIT is indicative of . If viable, this line of argument would raise problems for my second objection, insofar as it suggests that IIT's central claim is in fact correlational and thus, fails to falsify premise one of the neutrality argument[56]. However, this line of objection seems to be unpropitious for two broad reasons. First, a number of recent *philosophical* treatments of IIT suggest that the theory *does* commit its proponents to specific ontological positions and as such, cannot be read or interpreted correlationally in the manner just described. For example, it has been argued both that the current notion of information IIT utilises commits the theory to a physicalist metaphysics (Mindt 2017) and that, as it currently stands, the theory is currently incompatible with standard forms of Russellian panpsychism (Morch 2018)[57]. While the specifics of the ontological implications of IIT may be disputable, the claim that these recent treatments suggest - namely, that IIT *has* such constraining ontological commitments (and that these look initially to support a physicalist metaphysical thesis) - is sufficient to rule out the correlational objection to my argument above, and support the falsity of premise one of the neutrality argument. More broadly, the problem with this correlational objection to my argument is that in this context, the claim that IIT provides at best mere correlations seems to primarily stipulative, and reflects what I see as a wider failure in metaphysical discussions of consciousness to accept and take seriously the *explanatory* aims and

[56]See Chalmers' (2016), for the claim that we can view IIT as asserting an *a posteriori law of combination* as opposed to identity.

[57] Morch subsequently argues nonetheless that IIT as a theory should be subsequently revised to render the two compatible. Note however, that this latter sort of argument should not be attractive to the neutrality shipper, who requires that *current* neuroscientific theories, as opposed to those substantially revised - for possible ad hoc purposes - are compatible with the relevant competing metaphysical frameworks.

commitments at the heart of contemporary neuroscientific practice. I take up this claim below and in the following chapter.

3 Taking Stock

In this chapter I have argued that the prevalent assumption in recent metaphysics of consciousness that neuroscientific work is *metaphysically neutral or unilluminating,* which justifies metaphysicians working in methodological independence from neuroscientific work on consciousness, can be traced back to a single line of argument proposed by David Chalmers and more recently by Uriah Kriegel, which starts from a claim concerning the types of facts pursued by the neural correlates of consciousness research programme. I have argued that this 'neutrality argument' faces two serious objections: (i) that we have good reason to think that the correlational analysis of the NCC programme that this argument relies upon is mistaken and (ii) that, even so, the correlational work pursued by the NCC is no longer exhaustive of research in the field. These give us reason to think not only that the argument in its current formulation *fails to establish* the neutrality assumption, but also - given the strict ontological implications these objections suggest - that it is unlikely that a new neutrality argument can be reconstructed along similar lines.

What does the failure of the neutrality argument, as argued for here, mean for future methodological approaches to the metaphysics of consciousness? The proper conclusion of this chapter - and the claim I take myself to have argued for so far - is that the neutrality argument fails to establish the neutrality assumption, and as such - insofar there is no other readily available case in support of it - the robust methodological independence practiced in current debate, as outlined in the introductory chapter, lacks adequate justification. This should, I think, motivate us to consider and take seriously methodological alternatives to robust independence. I take this task up in the following chapters. Before doing so however, it is important to set out and summarise the main issue that lies behind much of the discussion in this chapter, so that the next step in the search for alternative methodological approaches (and the desiderata on such) is clear. Whilst the two objections presented here are distinct, they are both, I think, indicative of a related and larger problem which has rendered these previous arguments from Chalmers and Kriegel inadequate. This is that previous discussions have failed, to their detriment, to pay attention to the *specific theoretical and metaphysical commitments* of contemporary neuroscience. This is evidenced both in the correlational reading of the NCC programme which these arguments rely upon (which, as I and others have suggested, runs contrary to the prevalent form of explanation in neuroscience) along with the failure to take the more recent theoretical frameworks in consciousness science like IIT, along with their ontological implications, into serious consideration.

In order to establish the sort of metaphysical neutrality sufficient to justify robust methodological independence, it needs to be the case that the considerable work produced as part of the neuroscience of consciousness is devoid of ontologically constraining commitments. Another way of stating the main claims of this chapter (and the thesis which I examine further in the following chapters) is that, whilst this may have appeared plausible twenty years ago, recent work in the neuroscience of consciousness, along with parallel developments in the philosophy and metaphysics of neuroscience, suggest that this is increasingly implausible.

Chapter 3:
Towards a Neuroscience-First Metaphysics of Mind

"There are places where sophisticated scientific theses will cut directly against metaphysical ones, especially if the scientific thesis in question fits with established theory or enjoys indirect empirical support. Here there is danger for the scientifically näive metaphysician, and metaphysically informed work in general philosophy of science plays an important role in the refinement and development of metaphysical theories that involve such assumptions".

(L.A.Paul 2012;9).

In the conclusion of the previous chapter I argued that Chalmers' and Kriegel's neutrality argument failed to establish the neutrality assumption, and by extension, failed to provide justification for the methodological independence of the metaphysics of consciousness from the recent empirical work on its neural basis. I argued that this ought to motivate us to consider alternative methodological approaches to the metaphysics of mind, and furthermore, that this alternative approach should aim to accommodate - as a central desideratum - the sorts of theoretical and metaphysical commitments internal to neuroscientific practice which the arguments discussed in the previous chapter were suggestive of. In this third chapter, I present one such alternative approach. First however, I want to step back and provide this approach with a sound methodological basis. This takes as its starting point the recognition that the question at the heart of this thesis - namely, that of the correct account of the relationship between the metaphysics and neuroscience of consciousness - is an instance of the significantly broader *meta-metaphysical* question of the relationship between metaphysics and science more generally. This latter question has been the topic of heated debate in metaphysics over the past twenty years since Chalmers first put forward the neutrality argument. The starting point in my search and motivation of alternative approaches is thus the claim that, given its relationship to this broader question, the relationship between the metaphysics and neuroscience of consciousness ought to be sensitive to and illuminated by the key developments and argumentative insights within this field.

The plan for the chapter is as follows. In sections 1 and 2 I introduce and detail two developments from recent meta-metaphysics. These are (i) the emergence of a growing research project that attempts to detail and unpack the ontological and commitments of neuroscience and its explanatory practices and (ii) the prominent methodological defense of traditional metaphysics in response to the more radical discontinuation arguments in recent scientific metaphysics which claims that theory choice in metaphysics, as in some areas of theoretical science, proceeds via *inference to the best explanation*. In section 3 I combine and use these dialectical developments to construct a case in favour of a novel methodological approach to debates over consciousness' ontological status. This takes as central the claim that an approach to the metaphysics of mind which takes as its starting point the metaphysical commitments of neuroscience is not only an interesting potential alternative to methodological independence, but follows from reflection on the minimal *demand* that candidate ontological accounts of consciousness be *empirically equivalent*. In section 4, I outline the alternative 'Neuroscience-First' Approach to the metaphysics of consciousness which this argument suggests. I conclude by briefly discussing how this approach can be situated with respect to the aforementioned debates concerning so-called naturalised metaphysics.

1 The Metaphysics Of Neuroscience

A key theme in contemporary meta-metaphysics - the branch of metaphysics concerned with the nature and methodology of metaphysical inquiry - concerns the relationship between metaphysics and science. This debate is not new, having begun in earnest in Quine's arguments with the logical positivists in the 1940s, however it has had a contemporary resurgence in the programme of so-called "naturalised' or "scientific metaphysics". Here, the primary aim is to develop an account of metaphysical inquiry, and in most cases a resulting ontology, which is properly grounded in contemporary science. As stated on the website for the Society of the Metaphysics of science, the metaphysics of science programme takes as its broad remit "the abstract examination of ontological issues as they arise within, or grow out of, the sciences and their findings, concepts models or theories"[58].

This broad mandate has subsequently given rise to two distinct research programmes (Guay and Pradeau, 2017), the first of which attempts to utilise and apply recent empirical findings to augment areas of traditional metaphysical interest. This line of work, which tends to include contemporary debates over causality, laws and individuality (along with others) contrasts with the second programme - 'scientific metaphysics' - which starts from the claim that metaphysical inquiry is epistemically valuable *only insofar as it is grounded in* scientific observation and theorising and, as such, aims to construct an ontology based exclusively in contemporary (typically fundamental) science[59]. Within this latter, and arguably more prominent programme, the contemporary dialectic tends to proceed via by the construction of consecutive negative and positive arguments. Here, metaphysicians of science start by presenting a negative case a against traditional 'neo-scholastic' varieties of metaphysics - where the critical focus is typically on highlighting the various *epistemic inadequacies* of its a priori methods (such as its heavy reliance on intuitions, and minimal engagement with superficial characterisations of contemporary science) - which are subsequently used to motivate the construction of alternative *naturalised* metaphysical research programmes which, proponents claim, ought to serve as a replacement for traditional metaphysics.

Both the nature of the negative charges against traditional styles of metaphysics, and the extent to which these traditional programmes require replacement or discontinuation are matters of ongoing dispute in contemporary debate[60]. My primary concern here however is to highlight a broad development which has occured as a result

[58] https://sites.google.com/site/socmetsci/what-is-the-metaphysics-of-science-1.

[59] For examples of the former line of research, which I do not discuss further here, see Ellis (2001), Bird (2007), Lowe (2006), Chakravartty (2007) and Mumford and Tugby (2013). For scientific metaphysics, see Ladyman and Ross (2007), Maudlin (2007); Ross et al. (2013), Ney (2012), Wilson (2006) and Maddy (2007). Informative overviews of the field as a whole can be found in Tahko (2015;chp9) and Soto (2015) along with the essays included in Ross et al (2013).

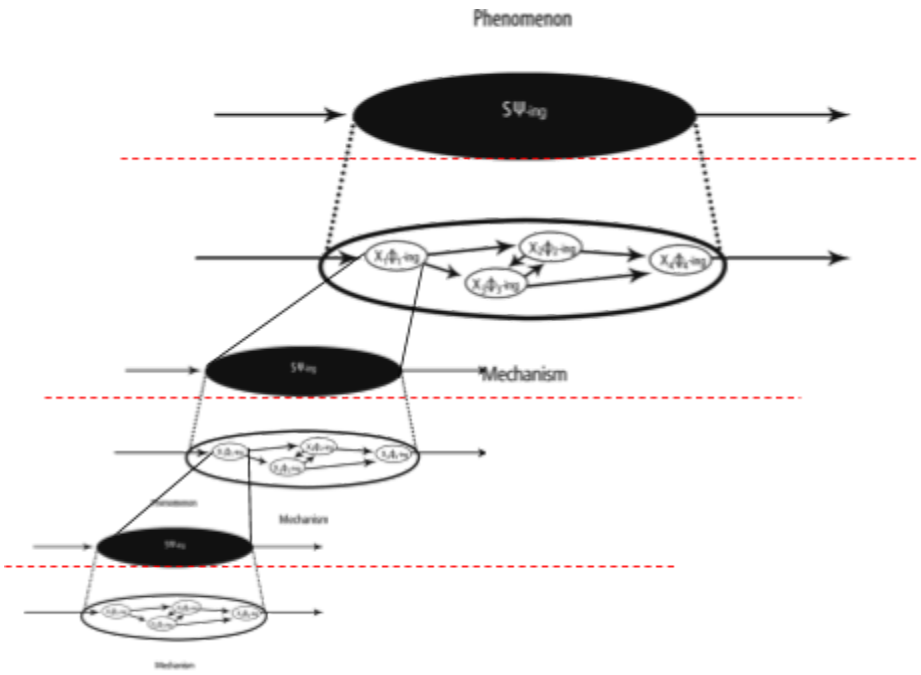
[60] For recent discussion of the purported epistemics sins of neo-scholastic metaphysics see Bryant (2017), Chakravartty (2013) along with French and McKenzie (2011). The nature of the metaphysical project which ought to replace traditional

of the growing interest in the programme of scientific metaphysics, broadly understood, and adherence to the wide research mandate highlighted in the quote above. This is the extension of the research programme, which (given the overlap in subject matter) in its initial formulations tended to focus on issues within fundamental physics, to the *life sciences*, and in particular, its recent manifestation in a growing body of work concerned with the metaphysical issues arising within neuroscience and neuroscientific explanation. That is, as result of the success of recent work in scientific metaphysics, a number of philosophers have sought to construct a richer and more inclusive programme of scientific metaphysics which, in addition to the traditional projects in the philosophy and metaphysics of physics, is expanded to include metaphysically oriented work in the philosophy of biology, psychology and neuroscience[61]. Conducted for the most part by the New Mechanists - the philosophers of science whose work I introduced in the previous chapter - the focus thus far in the metaphysics of neuroscience has primarily been on the development of a detailed account of the *metaphysics of mechanistic explanation*. That is, the aim of this recent work has been to provide of an account of the various ontological commitments which fall out of the detailed, descriptively adequate account of neuroscientific explanation detailed by contemporary philosophy of neuroscience, and the application of these to relevant debates in the philosophy of mind and philosophy of science[62]. Recall the account of explanation discussed in the previous chapter, depicted in the figure below:

styles of metaphysics - that is, the extent to which metaphysics requires *naturalisation* - is heavily disputed among scientific metaphysicians. Here, proposals range from the more radical viz. that metaphysics must be based solely on, and motivated by the sole purpose of serving or unifying, current science (Ladyman and Ross 2007, Ney 2012), to the more moderate. According to these more recent suggestions (Tahko and Morganti; 2017 Guay & Pradeu; 2017) whilst epistemic considerations demand that metaphysicians *must* engage consistently with contemporary science to a great degree, metaphysics can nevertheless retain its status as an autonomous discipline with a broad set of distinctive theoretical aims. This is defended either on the basis that metaphysics has a distinctive subject matter (see below) or that metaphysicians utilise a unique methodology (Tahko and Morganti 2017).

[61]For an explicit defense of a more inclusive metaphysics of science extended to the special sciences see Guay & Pradeu (2017), however all recent proposals for 'moderate' programmes of naturalised metaphysics include this suggestion.

[62]Descriptive adequacy here refers to accounts of explanation in neuroscience which is properly constrained by how neuroscience *actually* works and why it is successful.



Work in this emerging field is extensive, however prominent questions and issues discussed thus far include (i), the proper account of the *metaphysical constitution* of the mechanistic explanans - namely, what, metaphysically speaking, mechanisms are, and which ontological categories are best used to describe them and their component parts, (ii), an account of the ontological status of the *phenomena* which serves as the explanandum of mechanistic explanation - viz. what is the ontological nature of constitutive mechanistic phenomena, and as such, are they multiply realisable? (iii), what is the nature and relata of the mechanistic *level relation* (the dotted red lines) - and what does this mean for more traditional debates regarding *ontological reduction,* emergence and interdisciplinary unity in the special sciences? Finally, and most prominently, much of current discussion has concerned an issue which was touched on in the previous chapter, namely (iv) the correct account of the metaphysical *relationship* (the dotted black lines) which obtains between the mechanism and phenomenon to be explained in constitutive mechanistic explanation[63].

[63]Craver & Tabery (2017;S4), Glennan (2017; particularly chapters 6&7, and 2009), Krickel (2017, 2018), Baumgartner et al. (2018), Kaiser & Krickel (2016), Kastner (2017; Chps 7 & 11), Soom (2012), Harbecke (2010, 2014), Krickel (2014), Kaiser and Craver (2013), Craver (2007;chp5), Erononen (2013, 2015) and Gillet (2013). This metaphysical project seems prima facie to require that mechanistic explanation in neuroscience is fully, or at least partly *ontic,* as opposed to purely epistemic, in nature - that is, that mechanistic explanations are objective features of the world (Craver 2007;27, 2014, Salmon 1984, Illari 2013). This has not gone uncontested (see, for example, Glennan 2005, Bechtel and Abrahamsen 2005 Wright 2012, 2015), and the relationship between these alternative epistemic accounts of mechanistic explanation and the metaphysics of neuroscience deserves further consideration. Here and in the rest of thesis, however, I assume the validity of a (at least partly) ontic account of mechanistic explanation.

The details of the metaphysical picture of neuroscientific explanation which is emerging as a result of these discussions will be important later. For now, the purpose of my introducing this work, and the crucial claims I wish to make clear going forward, are threefold. First, that this emerging work heavily suggests that such a metaphysical picture of neuroscience *exists*. This adds further weight to the claim that a metaphysically 'vacuous' view of neuroscience and of neuroscientific practice discussed in the previous chapter is likely to be incorrect[64]. Second, that whilst this work on the metaphysics of neuroscience has emerged as a *consequence* and development of the metaphysics of science programme, its uptake is not parasitic on acceptance of the discontinuation thesis at the heart of much of recent scientific metaphysics. That is just to say, one can - if one rejects radical formulations of scientific metaphysics - take the research programme just outlined in the metaphysics of neuroscience to be both important and ontologically illuminating without endorsing the further claim that all of metaphysics should be reduced, or limited to, this and similar projects. Finally, it is important to note that the ontological commitments of neuroscience emerging from this branch of the metaphysics of science are the products neither of uncareful metaphysical speculation from those unfamiliar with metaphysics - as, it could be argued, is the case for the claims made by neuroscientists working on consciousness discussed in the last chapter - nor of detailed philosophical or metaphysical examination which floats free and is independent from the demands of actual scientific theorising. This latter point is crucial, and follows as result of the aforementioned commitment of the New Mechanists to the production of an account of scientific explanation which is *descriptively adequate;* that is, constrained and motivated by actual, as opposed to idealised, experimental practices used in neuroscience to successfully explain phenomena of scientific interest (Craver 2007;21, Kaiser and Krickel 2016;). This careful balance struck in the emerging work on the metaphysics of neuroscience between the cautious unpacking of the metaphysical structure of mechanistic explanations on one hand, and actual experimental practices on the other, gives us reason to take the emerging metaphysical picture very seriously; as that which is entailed by any serious theoretical acceptance of neuroscience, broadly understood, and its explanatory commitments.

2 Methodological Naturalism

The first development in metametaphysics which I take to be of direct relevance to the search for alternatives to methodological independence in the metaphysics of mind then is the recent work in the metaphysics of neuroscience. The second development I take to be crucial to making progress on this issue comes from a

[64]That is, this supports the view that the claim discussed in the Chapter 2 conclusion - that neuroscientific theories and practices entail specific ontological commitments - is not a unique feature of the neuroscience of consciousness, nor the naive claims made by its practitioners, but rather emerges as a plausible and *expected* consequence of recent philosophy and metaphysics of science.

different dialectical direction. Emerging in response to the more radical lines of arguments in recent scientific metaphysics just mentioned, which call the discontinuation and replacement of non-naturalised contemporary metaphysics - on the basis that its methodology, unlike that of science, is inadequate for the purposes of objective inquiry - philosophers have looked to defend traditional styles of metaphysics by providing an analysis on which metaphysics and science are broadly methodologically continuous. That is, according to proponents of *methodological naturalism*, theory choice in metaphysics proceeds in a broadly similar manner as theory choice in certain areas of highly theoretical science[65]. In both cases, proponents of methodological naturalism claim, the methods used to develop and subsequently select theories are the same.

The (much over simplified) story is as follows. First, a posteriori reasoning is used in the usual manner to develop competing theories of the world which fit the available observational data. Second, theory choice between these competing underdetermined alternatives is subsequently made on theoretical or a priori grounds, via an *inference to the best explanation* (IBE) (Paul 2012). That is, in the rare cases in which competing scientific theories are underdetermined by the empirical evidence, it is claimed that theory choice is subsequently warranted on the basis of an assessment of a theory's overall *explanatory power;* where this is understood in terms of a given theory's degree or maximisation of the evidential so-called 'super empirical' or theoretical virtues, including parsimony, plausibility, fecundity, internal consistency, universal coherence, and so on. The choice between Einstien's special relativity and Lorentz-type ether theories is typically offered as a case in point[66] .

[65]See Papineau (2015) for a discussion of methodological naturalism in a broader context, as distinct but related to naturalism understood as an ontological thesis. It is also important to distinguish the methodological naturalists' thesis from the physicalist *attitude* or stance one adopts when forming an ontology centered or based solely on current physics (Ney 2008, Van Fraassen 2002). Whilst acceptance of methodological naturalism may be necessary for endorsing so-called attitudinal varieties of physicalism (although this is debatable), it is not sufficient. That is, one can take metaphysics and science to be methodologically continuous in the manner described below, without - as we shall see - endorsing the further claim that the ontology produced via such methods is, or even ought to be, exclusively based on and limited to the ontological posits of contemporary physics.

[66] Ladyman (2012), although examples from other fields can be found in Tulodziecki (2013),Werndl (2013) and Belot (2014).This latter claim is usually expressed via the assertion that empirical equivalence is not exhaustive of the epistemic constraints on theory choice in science; that is, by the denial that the empirical adequacy of theories entails their *epistemic* equivalence, and as such, that theory choice in these cases is genuinely evidentially underdetermined. Here, the methodological naturalist enters into controversial ground. The epistemic or truth-conducive, as opposed to pragmatic, nature of the theoretical virtues employed in IBE has long been disputed in contemporary philosophy of science and, given its role in the the underdetermination argument against scientific realism, hotly debated (Van Fraassen 1980, Psillos 1999, Laudan and Leplin 1991, Tulodziecki 2012, Alai *forthcoming*). I do not wish to get in to these arguments here and as such, grant the viability of an IBE approach conditionally for the purposes of my argument. (C.f Ladyman's (2012) discussion of the relationship between this explanationist defense of metaphysics and the rise of scientific realism).

The methodological naturalists' claim then, is that metaphysical theorising proceeds analogously to these cases. In metaphysics, it is claimed, coherent metaphysical theories of a given phenomenon or feature of the world are first constructed, and subsequently integrated with philosophical logic. Whilst these may lack direct testability which is paradigmatic of most scientific cases, they must, like the theories just mentioned, nevertheless account for all available observational data with respect to the phenomenon and its observed instances. Given that there will be many such theories which meet these criteria (viz. coherence, logical consistency and empirical fit) the problem of choosing which theory to accept is solved in the same manner - by conducting a *cost benefit analysis* of each theory, where the relevant costs and benefits are thought to mirror the kinds of theoretical virtues utilised in theory choice between underdetermined alternatives in theoretical science.

Further discussion of this so-called 'explanationist' defense of metaphysics can be found in Chakraverty (2013), Nolan (2015) and Manley (2009) Ladyman (2012), however its main formulation, and that which has garnered most attention from its opponents, is L.A.Paul's analysis (2012). Crucial to Paul's particular explanationist defense of metaphysics is the claim that the shared a posteriori elements of scientific and metaphysical theorising are best understood in terms of *modelling* - where this is meant in the technical sense shared by proponents of the dominant semantic view of theories in contemporary philosophy of science, which equates theories with classes of models[67]. Whilst this is an important and interesting part of Paul's case for methodological naturalism and metaphysics as IBE, as Paul herself notes, adoption of the semantic view of scientific theorising is not necessary for her broader argument (12), nor is it relevant to the main line of argument I put forward in this chapter (see S3.2)[68].

Methodological naturalism as outlined allows for a response to the discontinuation arguments against traditional metaphysics which goes as follows. The central negative claim made by scientific metaphysicians is that traditional neo-scholastic style metaphysics, in contrast to scientific theorising, is unsuited for the purposes of generating substantial conclusions about the nature of the world in virtue of its *distinctive* a priori methodology (which they claim is epistemically defective). If methodological naturalism is true however, and metaphysics is correctly analysed in terms of inference to the best explanation, then this central claim of the negative discontinuation arguments is false. If theory-choice on this basis is warranted in science and everyday reasoning - which its proponents, along with scientific realists (as explained above), presume - so too, the methodological naturalist claims, IBE must be warranted in metaphysics. That is to say, on this view

[67]French & Ladyman (1999), Suppes (2002), Thomson-Jones (1997), Godfrey-Smith (2006).
[68]As such, I remain neutral on the semantic/syntactic debate, and do incorporate Paul's semantic account of metaphysical theorising into my discussion here.

metaphysics and science differ not - contra the claims of scientific metaphysicians - in the methodological practices they use to make claims about a subject of shared interest namely, the nature of objective reality, but rather, are best understood as engaging in and employing the same methodology with respect to *different* sorts of questions arising from the study of reality. As such, it is concluded, "those who argue that metaphysics uses a problematic methodology to make claims about the subjects better covered by natural science get the situation exactly the wrong way around" (L.A.Paul; 2012;1)[69].

The defense of metaphysics that an IBE or explanationist account of metaphysics allows for provides defenders of neo-scholastic metaphysics with strong reason to take this view seriously. Whether it is ultimately successful against the discontinuation arguments, however, remains to be seen, and will depend on the stances one takes on a number of further issues in the metaphysics and philosophy of science. Foremost, this will depend on the substantive issue of the legitimacy of IBE in science (and the outcome of related debates on the threat to scientific realism from underdetermination) mentioned above. Further, even if this is granted, the success of the explantionist defense of metaphysics just described will depend on whether one takes the rare cases of IBE and theory underdetermination in science to be analogous in the relevant respects so as to justify its widespread application in metaphysics. This latter claim viz. that the explantionist defense fails because the cases of theory underdetermination in science and metaphysics are *disanalogous,* is defended extensively in Ladyman (2012). Here, Ladyman argues that scientific theory underdetermination differs in a problematic sense from underdetermination cases in traditional metaphysics in two key respects, being both 'local' (domain or theory specific) and 'weak' (empirically equivalent in respect to *current*, as opposed to all potential, observations) (42-44)[70].

These issues deserve further consideration, however as before, it is not yet necessary to enter these broader debates concerning naturalised metaphysics. For now, my aim - beyond pointing out the primary dialectical motivation for methodological naturalism - is to reflect and make clear what this popular account of traditional metaphysics, *if* viable, commits us to with respect to the demands on metaphysical theories which are

[69]In Paul's view, metaphysical questions can generally be distinguished from those pursued by science on the basis of their generality and ontological priority (pp4-9). Thus, her claim that "metaphysics is concerned to identify the *real* nature of the world while science is concerned to discover the range of instances of these natures" (2012;5 emphasis added cf. Tahko & Morganti p12). I shall assume that an IBE analysis of metaphysics generally commits one to this 'different subject matter' claim, although I do not think that much hangs on this issue.

[70]That is, to use the usual terminology, it is claimed that instances of theory underdetermination in science are always transitive as opposed to permanent in nature (Stanford 2006). I return to this claim in the context of metaphysical theories of consciousness briefly at the end of this chapter.

constructed on the basis of an IBE to engage with contemporary science. As explained above, the basic claims that the IBE or 'explanationist' defense of metaphysics makes are as follows.

Metaphysical theorising proceeds via two consecutive steps:

(1) The construction of coherent and logically consistent metaphysical theories (or models) of features of the world which fit the available data. *('a posteriori step')*.
(2) Theory choice between competing empirically equivalent theories is made via the cost-benefit of the theoretical virtues that different theories exhibit. *('a priori step').*

With respect to direct scientific engagement then, the demands that an IBE defense places on metaphysical theories follows as a result of step (1), which is epistemically prior to (2). This is that an IBE analysis of metaphysics demands that candidate theories be *empirically equivalent*, that is, be consistent with all the available observational data:

Empirical Equivalence: in order to secure epistemic warrant, candidate metaphysical theories (those subsequently subject to the cost benefit theoretical analysis in step (2)) - must be empirically equivalent.

This requirement, which entails a sort of broad *compatibility* or consistency of metaphysical theories with contemporary science, is a minimal one. That is to say, the IBE analysis of metaphysics puts the bar for the scientific accountability of plausible metaphysical theories very low, and fails to place constraints on the formation of metaphysical theories which are robust and overly demanding[71]. This is important. From a dialectical perspective, if it can be demonstrated that contemporary metaphysical theories which plausibly demand analysis in IBE terms do not in fact satisfy this minimal condition, this will be bad news for the defender of contemporary metaphysics. If we have good reason to think that metaphysical theories constructed via IBE not only fail to satisfy the more stringent demands placed on them by contemporary accounts of naturalised metaphysics, but also the minimal demands revealed by their own defense via IBE, then we will have strong grounds to reject the epistemic credentials of such theories, and call for their replacement.

3. Empirical Equivalence Revisited

3 1 An Explanationist Analysis of the Metaphysics of Mind

[71](Bryant 2017).

I began this chapter with the claim that the search for alternative, non-neutral conceptions of the relationship between the metaphysics and neuroscience of *consciousness*, in virtue of the relationship between this question and the broader issue of the relationship between metaphysics and science, ought to be illuminated by examination of recent developments in meta-metaphysics. Situating this discussion in the context of the recent programme of scientific metaphysics, I explained how this programme has recently been extended to include work on the metaphysical underpinnings of neuroscience and neuroscientific explanation, and has subsequently thrown methodological naturalist accounts of metaphysics - as a response to the more radical interpretations of this programme - into greater prominence. How then can these two claims be used to make progress in providing an account of the proper relationship between the metaphysics and neuroscience of consciousness? A detailed examination of the metaphysics of mind in the context of naturalised metaphysics has not yet been undertaken. When it comes to the negative campaigns against neo-scholastic style metaphysics, debates have tended to target and focus on the viability of more traditional topics covered in metaphysics such as the contemporary debates over universals, persistence, and causation. The prospects of naturalising 'applied' debates, such as those in the metaphysics of mind, have so far been left unexamined. Given the primacy of these debates in contemporary scientific metaphysics, this philosophical task is urgently required. The first positive claim that I wish to put forward is thus an attempt to begin this discussion, whilst aiming to construct an alternative approach to neutrality and methodological independence. This is the claim that the metaphysics of mind is best understood as falling within the methodological naturalist framework just outlined. That is, that the metaphysics of mind as it is currently practiced proceeds via inference to the best explanation, from candidate ontological theories of consciousness which are purportedly empirically equivalent.

I take this claim to have two broad motivations. The first should be obvious in light of the previous discussion. In addition to being both a prominent contemporary analysis of metaphysics independently of these debates, and seemingly implicit in many of the recent discussions of the metaphysics of mind mentioned in this thesis, an IBE account of the metaphysics of mind is motivated by the fact that it provides the metaphysician of mind with the resources to construct a case, a la Paul, against the discontinuation arguments at the heart of recent scientific metaphysics[72]. This is far from trivial. In the absence of compelling arguments against these claims -

[72] With respect to these earlier points, the IBE meta-metaphysical analysis of the metaphysics of consciousness is implicit in Balog's (*forthcoming*) diagnosis of the ontological stalemate (and in particular, her argument concerning the lack of a posteriori *and* a priori means of breaking this dialectical deadlock;18) as well as featuring prominently in Kriegel's (*forthcoming*) neutrality argument discussed at length in Chapter 2 (and his account of neutrality as empirically equivalency, more on this below). For discussion of the IBE analysis outside of the metaphysics of science see the discussion in Beebee (2018;3) - who, whilst not endorsing this meta-metaphysical position, nevertheless takes the IBE analysis to be the dominant or default position in contemporary meta-philosophy.

regarding which, I think it fair to say, there is no current consensus - these arguments have radical consequences for the future of metaphysics and its contemporary practice. Insofar as an IBE account of the meta-metaphysics of mind allows for a strong and potentially convincing response to the discontinuation arguments, as opposed to leaving contemporary metaphysical accounts of consciousness vulnerable to these epistemic charges, metaphysicians of mind ought to take this claim seriously.

The second motivation for the view is most important. This is the claim that an IBE analysis of the metaphysics of mind is plausible not only from a dialectical point of view in meta-metaphysics and the metaphysics of science, but also ought to be accepted on the basis that it fits and accounts for what metaphysicians of mind *actually do*; that is, to borrow the term from philosophy of science, it provides an analysis of the meta-metaphysics of consciousness which is *descriptively adequate*. Recall the broad outline of contemporary dialectic in the metaphysics of mind outlined in Chapter 1. There, it was explained that contemporary developments in debates over the ontological status of consciousness have been *both* a posteriori - take the debates following the acceptance of causal closure to be a case in point - as well as priori in nature, with the majority of discussion falling under this latter description.

This varied nature of contemporary debate plausibly serves as a central explanandum on any potential account of the meta-metaphysics of consciousness. However, it is one that an IBE analysis of the metaphysics of mind seems best placed to explain. On an IBE analysis of the metaphysics of mind, for example, we can make sense of the former a posteriori elements in terms of the attempt to demonstrate that various candidate theories of consciousness - namely, interactionist dualism - in light of recent empirical developments, no longer satisfy the empirical equivalence condition demanded by step (1) of the IBE analysis (with empirical evidence now pressing against it) and thus, are rejected as viable candidate hypotheses. Similarly, the attention and subsequent focus in contemporary debate on a priori arguments, both in the standard anti-physicalist or mental causation cases, along with the more recent naturalistic non-physicalist programme (which, recall, takes as a *holistic* treatment of non-physicalism beyond narrow consideration of the anti-physicalist arguments as its broad mandate) is explained by appeal to step (2) of an IBE framework. That is, the a priori arguments in metaphysics of mind following causal closure are plausibly understood, on this view, as contributing to the sophisticated cost-benefit analysis of competing metaphysical theories, modelled on the super-empirical virtues implicated in IBE in science. All of which, crucially, proceeds on the assumption that these candidate metaphysical theories satisfy Empirical Equivalence.

The two considerations above are not conclusive arguments for the claim that metaphysics of mind proceeds via inference to the best explanation. In the absence of compelling alternatives they do, however, provide strong motivation to accept such a claim. In the remainder of this thesis I assume that an IBE analysis of the metaphysics of consciousness, at least, restricted to the debate concerning consciousness' ontological status, is broadly correct. The following argument can thus be understood conditionally: *if* the metaphysics of mind is correctly analysed in terms of IBE, what follows with respect to the demands placed on contemporary metaphysical accounts of consciousness to engage with the recent empirical work on its neural basis? (and moreover, with respect to the ability of contemporary theories to meet such demands). Here, the next piece of the methodological puzzle - the recent work in the metaphysics of neuroscience - becomes key. This is because, as I will argue below, this metaphysically informed work in the philosophy of neuroscience, broadly understood, has significant implications for answering the question just posed, in virtue of providing a novel way of interpreting the demand for *empirical equivalence* of theories dictated by an IBE analysis of the metaphysics of mind. That is, when it comes to the satisfaction of empirical equivalence by competing *metaphysical* theories of consciousness, there may be a further or stronger sense of empirical equivalence suggested by the emergence of the metaphysics of neuroscience beyond that utilised in cases of theory underdetermination in science, which - I will argue - it is nevertheless reasonable to expect of candidate metaphysical hypotheses with respect to consciousness to satisfy[73].

3 2 Observational and Metaphysical Empirical Equivalence

If, as I have suggested, the construction and subsequent debate over competing accounts of consciousness' ontological status proceeds broadly via inference to the best explanation, the demand on contemporary

[73]Before continuing, it is worth briefly addressing two potential alternative accounts of the meta-metaphysics of mind suggested by literature in recent metaphysics of science. These are meta-metaphysical accounts of consciousness which either (i) reject methodological naturalism on the basis of the claim that, in addition to its distinctive subject matter, metaphysics is also autonomous from science with respect to its methodology - being, for example, *exclusively a priori* in nature (Lowe 2011;101), or, alternatively, (ii) endorse a deflationary instrumentalist view of the metaphysics of consciousness, which rejects the claim that metaphysical inquiry is capable of generating knowledge of the fundamental nature of objective reality, but nevertheless leaves room for its having various pragmatic or instrumental benefits (see for example, Beebee 2018, French and McKenzie 2011, Godfrey-Smith *unpublished*). This last alternative, suggested by the more recent discontinuation arguments, should be unattractive to metaphysicians of mind for obvious reasons. The second, whilst perhaps suited to some areas of contemporary metaphysics (see Tahko and Morganti; (2017;3) for an argument that Lowe's analysis is best placed to explain the methodology of metaphysicians engaged in conceptual analysis), seems ill suited as an account of the meta-metaphysics of consciousness for the reasons given above. That is, it is difficult to see how this analysis can account for the a posteriori developments central to the metaphysics of mind outlined above (and thus ought to be rejected on the basis of descriptively inadequacy) and provide a response to the prominent discontinuation arguments.

metaphysical theories to engage with neuroscientific research on consciousness - that is, the proper account of the relationship between these two lines of research - is easily identified. Simply stated, an explanationist analysis of metaphysics demands that competing ontological accounts of consciousness must satisfy Empirical Equivalence. In order to be considered an adequate metaphysical theory of consciousness, that is, whose costs and benefits, internal inconsistencies and viability etc. are worth debating in step (2) of the methodological process, a given ontological account of consciousness (non-reductive physicalism, Russellian monism etc.) must, as a minimal condition, be *consistent* with all available empirical data emerging from the new neuroscience of consciousness[74].

Given the lack of philosophical clarity on the question of the relationship between the metaphysics and neuroscience of consciousness that we began with in Chapter 1, the identification of this criterion via the explanationist defense, as a means of systematically addressing the empirical adequacy of competing metaphysical accounts of consciousness, constitutes good progress. However, this claim leaves open a crucial question which needs addressing if this criterion is to be properly informative viz. what does empirical equivalence amount to specifically, and how is the demand for empirical equivalence of competing theories best interpreted in *metaphysical* cases of theory underdetermination by empirical evidence?

The provision of an account of empirical equivalence in the standard scientific cases discussed above is itself a substantive issue, whose comprehensive treatment goes well beyond the scope of this thesis. Fortunately, there is a well accepted definition of the empirical equivalency had by rival theories in science which has emerged as a result of the recent debates over theory underdetermination and scientific realism, which will suffice for our purposes here (namely, getting a clear understanding of what is meant by empirical equivalence in the scientific cases of IBE). This can be stated as follows:

Observational Empirical Equivalence[75]: A set of rival theories (T_1, T_2, T_3...T_x) are empirically equivalent at time t iff (i) theoretical systems T_1..T_x have the same class of observational consequences at t or (ii) the set of theories has the same class of empirical models at t.

[74]Strictly speaking of course, rival ontological accounts of consciousness must be equivalent with respect to *all* relevant observational data, not just that produced by contemporary neuroscience, but I leave this here.
[75]Tułodziecki (2012;315), Worrall (2011), Laudan and Leplin (1991), Psillos (1999). Following Tułodziecki's assumption that a semantic or syntactic analysis makes no difference to assessments of empirical equivalence (on ease of translation from one to the other) I limit my focus in subsequent discussion to (i).

This observational account of empirical equivalence in science, which characterises empirical equivalence primarily in terms of the empirical *indistinguishability* or congruence of rival theories (Tulodziecki; 2012) serves as an obvious starting point for an account of empirical equivalence in the metaphysical case at hand. That is, in the first instance, we can interpret the demand on rival metaphysical theories of consciousness M_{T1} and M_{T2} to be empirically equivalent in terms of the demand on M_{T1} and M_{T2} to share *all the same observational consequences*. On this view then, we can claim that M_{T1} and M_{T2} are empirically equivalent iff for every empirical prediction e made by M_{T1}, M_{T2} also entails e[76]. Indexing the demand for empirical equivalence in this case to the neuroscience of consciousness, an observational interpretation of empirical equivalence yields the claim that metaphysical accounts of consciousness M_{T1} and M_{T2} are empirically equivalent with respect to this recent work iff they yield the same observational consequences in this context (that is, where the empirical prediction e above is one which is testable within this restricted empirical domain).

The observational empirical equivalence just described serves as a plausible interpretation of the demand on ontological accounts of consciousness to be empirically equivalent as dictated by step (1) of the explanationist defense of metaphysics. Presumably, no ontological account of consciousness would be accepted as a plausible candidate hypothesis concerning the metaphysical nature of consciousness, and subject to subsequent a priori cost-benefit analysis, if it was empirically inequivalent in this observational sense. The rapid decline in support for interactionist dualism due to its lack of observational confirmation from modern neuroscience stands as testament to this claim. Moreover, the demand for observational equivalence, as the operational account of empirical equivalence in scientific cases of IBE, must presumably be a demand which is shared by metaphysical theories constructed via IBE if the methodological naturalist defense of the metaphysics of mind is to be viable. As such, it is plausible then, that observational equivalence is *necessary* for the empirical equivalence of rival metaphysical accounts of consciousness. The question remains however, is it sufficient? That is, can the

[76](Worrall 2011). The observational account of empirical equivalence just outlined in fact tracks two distinct definitions of empirical equivalence offered in the context of theory underdetermination, which are distinguished by their different explication of the term 'observational consequences'. The first, 'naive' view, defines the relevant observational consequences of a set of rival theories in terms of results or predictions which are *directly testable*. On this view, two competing scientific theories (or rather, Duhemian theoretical systems) T1 and T2 are equivalent at t iff for every empirical prediction e made by T1, T2 also entails e - where e is a prediction capable of being *directly experimentally tested or checked* at t (Worrall 2011;162). This so-called "data equivalence" of rival theories (which I think captures much of what is intuitively meant by empirical equivalence in explanationist defenses of metaphysics as that which " share all the same empirical success") has since been deemed to be insufficient for genuine empirical equivalence (e.g. Worrall 2011, 2010) and has been subsequently replaced by a more stringent account of prediction e in the broad definition above in terms of the total observational consequences of a theory which are "expressed in a purely observational vocabulary" (Laudan and Leplin 1991). This latter account of equivalence is supposed to better accommodate, for example, the choice between Copernican and Ptolemaic models of the solar system which stand data equivalent, despite being empirically inequivalent. In my discussion here, I take this observational account to incorporate both of these more specific senses of equivalence.

observational account exhaustively capture the demand on ontological accounts of consciousness to be congruent and equivalent with respect to the large body of recent empirical work on its neural basis?

Before the emergence and widespread acknowledgement of scientific metaphysics as a fruitful research programme, such a claim might well have been plausible. However, the recent work in meta-metaphysics described in section 1 gives us strong reason to doubt the claim that satisfaction of observational equivalence by two rival ontological theories of consciousness is sufficient to establish their *empirical* equivalence. By all accounts, the metaphysics of science as a metaphysical research programme is revealing - in no uncertain terms - that the rich variety of explanatory theories and experimental practices in science, including those in neuroscience and related fields, are themselves *metaphysically loaded*, entailing or 'bringing with them' a set of precise metaphysical claims and commitments. Furthermore, this metaphysically informed work in the philosophy of neuroscience, and the metaphysical structure of neuroscientific explanations this reveals, is not detached or 'free floating' with respect to the relevant empirical results implicated in observational equivalence - so as to be *discretionary* with respect to acceptance of the results in these fields- but instead are tightly connected to them. That is, being, as described, constrained in the first instance directly by these practices via a methodological commitment to descriptive adequacy.

When it comes to the empirical equivalence of rival *metaphysical* theories of consciousness with respect to a given field then, there now appears to be a further condition beyond observational equivalence which demands satisfaction. This is the securement of the consistency or compatibility of these competing ontological theories with the best account of the rich metaphysical structure entailed by the theories and successful explanatory practices within this domain. In other words, the recent emergence of scientific metaphysics allows for a novel understanding of the demand for empirical equivalence on metaphysical theories of consciousness constructed via IBE, which can be stated as follows:

Metaphysical Empirical Equivalence:[77]. A set of rival metaphysical theories (MT_1, MT_2, MT_3...MT_x) are empirically equivalent with respect to domain y at t iff (i) they are observationally equivalent at t and (ii) they are consistent, or equally compatible, with our best account of the metaphysical commitments of y at t.

The demand for metaphysical, in addition to observational, equivalence on rival metaphysical theories, if viable, potentially sets the bar for the empirical adequacy of ontological accounts of consciousness significantly higher.

[77]See also Worrall's (2011) discussion of the necessity of various theoretical commitments in securing empirical equivalence.

This implication, I think, ought to be embraced by proponents of explanationist accounts of the metaphysics of consciousness. The establishing of the empirical adequacy of competing metaphysical theories in step (1) of IBE, prior to subsequent assessment of a theory's overall explanatory power, is crucial, and, as explained, is designed to secure the theory's complete empirical congruence or *indistinguishability*. If methodological naturalists are serious about meeting this minimal prior commitment (and, by extension, securing the proper continuity of scientific and philosophical knowledge in this case) the mere compatibility of metaphysical theories of consciousness with our best accounts of the *metaphysical commitments* of neuroscience ought reasonably to be secured[78].

In light of the above, the alternative methodological claim regarding the relationship between the contemporary metaphysics and neuroscience of consciousness I am proposing should be clear. This is straightforward: that close attention to, and reflection on, the methodological practices adhered to in the metaphysical debates on consciousness (IBE) reveals that the correct approach to the metaphysics of consciousness is one which looks in the first instance to secure the *metaphysical equivalence* of rival ontological accounts of consciousness (which has currently been ignored). That is to say, the right approach to formulating metaphysical accounts of consciousness and its ontological status is one which starts from close attention to the details of the emerging metaphysics of neuroscience, and specifically, of the neuroscience of consciousness, and utilises these to construct competing positive (empirically adequate) metaphysical accounts of consciousness - from which traditional-style a priori metaphysical theorising can begin. To state this using more familiar terminology, when attempting to answer the hard problem of consciousness (viz. 'why and how' a given set of physical processes should give rise to subjective experience), we should start from attention to the metaphysical commitments entailed by the solutions of the so-called 'easy problems' of consciousness, which are provided by cognitive neuroscience and related empirical fields. Here, the claim is not that attention to these empirically informed metaphysical commitments will necessarily *solve* the hard problem (and/or provide compelling answers to its related argumentative formulations) but rather, that the framework for answering the easy problems of consciousness collectively place empirically motivated *epistemically prior* constraints on the metaphysical space of its potential a priori solutions (see figure 2).

[78]This constraint, which is significantly more *robust* than observational equivalence, ought also to be attractive insofar as it provides the explanationist defender of the metaphysics of consciousness with the potential resources to combat a number of the recent epistemic arguments pressed against the methodological practices of traditional metaphysics (see below).

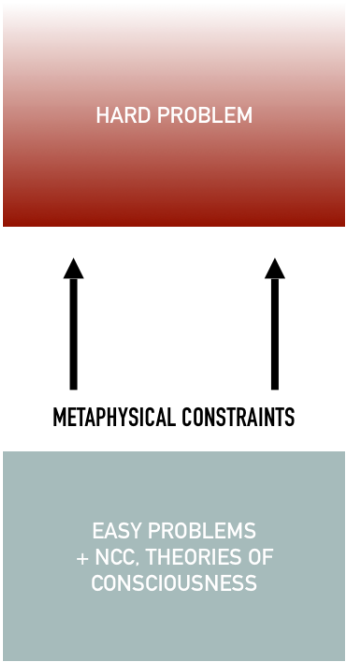


Figure 2: Easy vs Hard Problems and Metaphysical Equivalence

Before continuing onto the exposition of the alternative methodological approach which this discussion suggests, it is necessary here to briefly return to the arguments I presented in the previous chapter against the neutrality argument. The distinction between the two types of empirical equivalence made in this chapter allows for a restatement of the argument presented in Chapter 2, which reinforces those arguments and situates them within this broader methodological context. This can be stated as follows: that whilst Chalmers' and (to a greater extent) Kriegel look to, or attempt to secure the *observational* equivalence of competing ontological accounts of consciousness, the neutrality argument fails primarily on the basis that Chalmers and Kriegel fail to secure, or even address, the *metaphysical equivalence* of rival ontological accounts with the neuroscience of consciousness (which, given the preliminary discussions of causation and constitution in the mechanistic interpretation of the NCC, and the identity claim in IIT, looked difficult to establish). That is, from the discussion above we can see that Kriegel was *correct* in thinking that the neutrality of the neuroscience of consciousness would result from the claim that competing metaphysical theories are empirically equivalent with

respect to this field, but that he was nevertheless mistaken in assuming that we could infer the truth of this claim from the observational or data equivalence of one or two (or even all) such metaphysical theories[79].

4 A Neuroscience-First Approach: A Proposal

As explained above, the alternative methodological approach that is suggested by the discussion of recent developments in meta-metaphysics covered in this chapter takes as its broad mandate the securing (in the first instance) of the empirical - that is, metaphysical *and* observational - equivalence of competing metaphysical theories of consciousness within an explanationist or IBE framework. However, as should be evident from the different objections discussed in Chapter 2, there are various ways in which this broad mandate can be implemented. As I understand it, this alternative approach to the metaphysics of consciousness which looks to use the emerging work in the neuroscience of consciousness to secure the empirical equivalence of theories in this sense, is constituted by the pursuit of the two questions, considered in turn:

(1) What are the unified set of metaphysical claims or commitments emerging from careful philosophical examination of the neuroscience of consciousness?

(2) Which leading metaphysical theories of consciousness are consistent or compatible with this set of claims? (or are these instead suggestive of a single ontological picture?)

Together, these constitute the alternative 'Neuroscience-First' methodological approach to the metaphysics of mind. With respect to the first of these questions, it is important to distinguish two distinct approaches which can be taken, both of which are reflected in the two objections to the neutrality argument discussed in the previous Chapter. First and most obviously in light of the discussion in this chapter, one can work towards identification of the metaphysical commitments of the relevant empirical work by adopting a *discipline-specific* approach. That is, an approach which takes as its starting point the ontological commitments of the (in this case, mechanistic) explanatory practices characteristic of neuroscience and related fields, and applies these, where appropriate, to relevant work in the neuroscience of consciousness. An example of this approach in relation to the NCC framework is provided in the next Chapter. As already explained, however, the contemporary research produced as part of this NCC framework (to which the discipline specific approach looks to be best suited) is not exhaustive of neuroscientific work on consciousness, which is also comprised of

[79]As I argued in 2.2 and 2.3, I think that as things currently stand, Kriegel fails to establish even the observational or data equivalence of all of the standard metaphysical theories, insofar as he only examines the observational equivalence of two of these candidate hypotheses (viz. constitutive non-reductive physicalism and epiphenomenalism).

significant work relating to and testing comprehensive empirical theories or models of consciousness. An alternative and distinct *theory-approach* to (1) then, attempts to answer this question by drawing out the main metaphysical commitments of these leading theories as they emerge from the science of consciousness[80]. A complete answer to (1) will thus require the integration of discipline and theory-specific approaches, however - as we shall see in the next chapter - there is plenty of work to be done separately within the different approaches before this can be done.

5 The Metaphysics of Consciousness, Naturalised?

The Neuroscience-First Approach just outlined constitutes a significant departure from the usual methodological approach in the metaphysics of consciousness, according to which metaphysical and neuroscientific research operate in robust methodological independence from one another. On this alternative view, contemporary empirical research in the neuroscience of consciousness and its metaphysical foundations are not independent from metaphysical theorising on consciousness but rather play an indispensable role within it, serving as the mutual starting constraints on rival empirically adequate ontological theories of consciousness. In this chapter, I have argued primarily that this approach follows from two claims which have gained plausibility in light of recent developments in meta-metaphysics (i) that the metaphysics of mind proceeds broadly via inference to the best explanation from empirical equivalent theories and (ii) that empirical equivalence in metaphysical cases of IBE demands not only observational or data equivalence of rival theories, but also congruence with our best (carefully extracted) metaphysical theories of neuroscience.

As should be evident from my introduction and explanation of (i) and (ii), both of these claims used to motivate this alternative methodological approach have emerged as a result of the broader debate in meta-metaphysics which concerns the extent to which traditional styles of metaphysics and its methodological practices (of which, metaphysics of mind is presumably a part of) are epistemically inadequate and require replacement. One of the obvious questions left over from this discussion is thus where a Neuroscience-First Approach to the metaphysics of consciousness (or one like it) leaves the metaphysics of mind in the context of such debates. A

[80]As such, this would include work like the recent metaphysical treatments of IIT discussed in the last chapter (Mindt 2017, Morch 2017), along with metaphysically informed discussions of the PEM frameworks, global workspace theories and (if and when this emerges) the comprehensive theoretical framework in consciousness science which unifies these disparate frameworks. I also take this theory-specific approach to incorporate the emerging work on the metaphysics of global states or levels of consciousness (Bayne & Howhy 2016) and the recent empirical work proposed as part of the meta-problem research programme (Chalmers forthcoming). This fits nicely into this methodological Neuroscience-First programme insofar it is committed to the claim this empirical work ought to to 'shed light on it's[the hard problem] possible solutions'.

further benefit of the Neuroscience-First approach, and a *prima facie* reason to take the view seriously in this context beyond the argument presented here, is as follows. That, whilst this approach relies on and utilises an explanationist analysis of metaphysics, it nonetheless has the resources to satisfy a number of recent criteria proposed by scientific metaphysicians for epistemically adequate metaphysical theorising. That is, this alternative methodological approach places empirically motivated constraints on possible theories of consciousness which are *robust* and demanding (Bryant 2017) (such that the number of potential metaphysical theories of consciousness will be reduced, and epistemic warrant secured) and also, allows for instances of theory underdetermination in metaphysics of mind which are both *transient* in nature (that is, broken by new data) and specific to given domain (neuroscience) and thus, on certain accounts, properly analogous to purported cases of empirical underdetermination in science (Ladyman 2012). In other words, the Neuroscience-First Approach outlined here has further benefits as a methodological approach to the metaphysics of consciousness insofar as it provides a *prima facie* example - pending closer philosophical examination - of how the metaphysics of mind can proceed within a moderate programme of naturalistic metaphysics (Guay and Pradeu 2017, Tahko and Morganti 2017).

Chapter 4:
An Application: The Neural Mechanisms of Consciousness

In the previous chapter I argued that the standard methodological approach to the metaphysics of consciousness ought to be replaced by one which begins with the identification and examination of the metaphysical commitments of emerging work on its neural basis. Whilst this methodological debate is philosophically interesting in its own right, the proper measure of an alternative methodology is surely in its usefulness and application to contemporary debates. My aim in this fourth chapter is thus to demonstrate how this alternative Neuroscience-First approach can be implemented in current discussions, and can be used to constructively inform ongoing debates in the metaphysics of mind. Thus far, the target and focus of my discussion has been exclusively on a series of methodological claims concerning contemporary theorising in the metaphysics of consciousness, and how this ought to proceed; the viability and details of the various rival metaphysical accounts which have emerged as a result of the standard methodological procedures - that which the majority of contemporary work in the metaphysics of consciousness is concerned with - has been left undiscussed. This has been deliberate, and reflects the prior objective to provide a detailed and constructive discussion of the methodology and meta-metaphysics of consciousness which is free from metaphysical ideology and first-order agendas[81]. Here however, the aim is to move away from this focus on methodology and begin to consider some of the implications this Neuroscience-First approach might have for specific first-order concerns in recent metaphysics of mind.

The plan for this final chapter is as follows. In **Section 1** I return to the recent mechanistic interpretation of the Neural Correlates of Consciousness research programme which was introduced in Chapter 2. As this is paradigmatic of the discipline-specific approach outlined in the previous chapter which draws directly on recent work in the metaphysics of neuroscientific explanation, the mechanistic interpretation of the NCC programme and its metaphysical structure serve as a natural starting point for the implementation of the Neuroscience-First Approach. After briefly recapping the arguments and discussion presented in 2.2.1, I outline the main claim which emerges from this mechanistic interpretation of the search for NCCs. Whilst I am in broad agreement with the motivation and interpretation of the programme offered by Neisser (2012), the account presented here differs from Neisser's in details, and amounts to the claim that the search for neural correlates of consciousness is best interpreted on a mechanistic framework not in terms of mechanistic causes, but as the search for the multi-leveled neural mechanisms *which are constitutive* of consciousness. That is, I argue that Neisser's mechanistic definition of the NCC programme is currently inadequate as it mistakenly conflates constitutive with etiological forms of mechanistic explanation, which is unsuited to this experimental context (Craver 2007;74).

[81]As such, I have tried to suspend judgement on the familiar first-order issues whilst considering these methodological questions.

If this constitutive mechanistic analysis of the search for the Neural Correlates of Consciousness is broadly correct, we should expect this explanatory mandate to be reflected in the common experimental paradigms utilised by researchers working in this research programme. In **Section 2,** building on Neisser's previous discussion of Binocular Rivalry studies, I argue that the standard experimental paradigms used as part as the contemporary search for the NCCs support this particular constitutive-mechanistic analysis. Taking the recent report and 'no-report' based paradigms in content-based studies as a key example (Aru et al. 2012 Tsuchiya et al. 2015), I argue that these and similar experimental approaches are best understood as instances of so-called 'Activation Experiments' (Craver 2007, Kastner, 2017) viz. a specific type of top-down interlevel experiment commonly used to identify *constitutive components* of a given neural mechanism. After having provided this further motivation for the mechanistic analysis of the NCC programme, in **Section 3** I argue that this provides the basis of a novel argument against contemporary type-identity views of consciousness. This takes as central the straightforward claim that ontological accounts which claim that consciousness is *identical* to some neural process or mechanism are in tension with the widely accepted view that the relationship between mechanism and phenomenon in constitutive mechanistic explanations are non-causal *constitutive* dependency relations. I conclude with a discussion of further work to be done as part of this mechanistic project applied to the neuroscience of consciousness.

1 Causal vs Constitutive Mechanistic Explanation

In Chapter 2, I argued that a mechanistic analysis of the NCC programme gives us reason to reject the claim that neuroscience, as it pertains to the study of consciousness, aims at the production of purley correlational facts. According to this mechanistic analysis (Neisser 2012, along with Revonsuo 2000, Vernazzani 2015 and Seth 2009), the dominant form of explanation in neuroscience, as detailed extensively by the New Mechanists, places interpretational demands on an account of the NCC programme, such that this is best characterised as the search for an *explanation* of consciousness via the description and identification of the multi-levelled *neural mechanisms* responsible for its production. When applied to both content-specific and stated-based approaches constitutive of the contemporary NCC programme, I argued that this mechanistic analysis supports the claim that researchers working in NCC programme are searching for facts relating first person and third person relata which are metaphysically specific and *non-correlational*, in a manner which would render the neutrality argument unsound[82].

[82] A detailed description of mechanistic explanation (in addition to the Neural Correlates of Consciousness research programme) has been provided earlier in the thesis. See Chapter 2 (2.2.1) and Chapter 1 (1.2.1), along with the references included there for more extensive treatments of these research programmes. As before, I limit my discussion of mechanistic explanation here primarily to the influential account offered in Craver (2007) (C.f.Bechtel 2005, 2007, Darden 2008,

But what account of this non-correlational relation obtaining between a given conscious content (or individual global states) and neural mechanisms does a mechanistic analysis of the NCC programme demand? In Chapter 2, I stopped short of endorsing a particular view of this relationship. As explained previously, in his mechanistic account of the NCC programme, Joseph Neisser (2012) proposes a *causal-mechanistic* analysis of the NCC programme. That is, Neisser argues that the type of facts pursued by NCC researchers are best characterised as *causal*; a claim which is thought to follow as a consequence of the broad mandate adopted by the New Mechanists to situate diverse cognitive phenomena within the *causal structure* of the world (Craver 2007). This causal-mechanistic analysis is summed up in the revised definition of an NCC that Neisser provides (which he restricts to content-specific NCC approaches) which goes as follows:

"An NCC can be defined as a minimal neural system N such that states of N are underlying **causes** of a measurable change in consciousness, where a given state of N, as the causally relevant component of an embodied mechanism, is a mutually manipulable **INUS** condition for the specified aspect of the conscious state" (2012, 689).

This causal-mechanistic analysis provided by Neisser was useful for the argument I presented in Chapter 2 (which aimed to demonstrate that a mechanistic analysis of the NCC programme ought to give us reason to reject the standard correlational reading of this empirical research) and provides a concrete example of how the NCC project might be analysed within a mechanistic explanatory framework. The causal-mechanistic analysis of the NCC project given above however is not the only analysis of this research that can be provided by attention to the details of recent literature on mechanistic explanation. According to an alternative analysis of the NCC research programme, alluded to in my previous discussion, contemporary researchers working in the NCC project are not looking to locate the mechanistic *causes* of a given content (or states) of consciousness, but rather are looking to identify the acting component-entities of the neural mechanisms which are *constitutive* of the relevant aspect of consciousness[83]. This distinction between competing causal-mechanistic and constitutive mechanistic analyses of the NCC project (which nevertheless share the same broad motivation) reflects the common distinction found in the new mechanist literature between etiological and *constitutive* forms of mechanistic explanation[84]. Whilst in etiological mechanistic explanations phenomena are explained by their prior mechanistic causes, in constitutive mechanistic explanations - which are arguably most prominent in neuroscience - higher level cognitive phenomena are explained instead by the identification and description of

Glennan 2005, 2017). A comprehensive discussion of the NCC programme and its recent experimental developments, which my discussion below targets, can be found in Koch et al. (2016) and Howhy & Bayne (2016).

[83] This is suggested in Revonsuo's comprehensive discussion of the NCC framework in 2000 (chp1-3).

[84]See Salmon (1984), Craver (2007, 65-72), Kaiser and Krickel (2016;7-9).

the neural mechanism which constitutes or 'underlies' it[85]. This latter form of explanation, which is most commonly discussed in the philosophy and metaphysics of neuroscience, was outlined and explained in Chapter 2. As a way of brief recap, its main claims can be summarised as follows (1-5):

1. That (neuro) scientific explanation proceeds by uniformly identifying and describing multilevel neural mechanisms that are responsible for diverse higher level cognitive phenomena. This typically involves the integration of numerous fields and empirical methodologies.

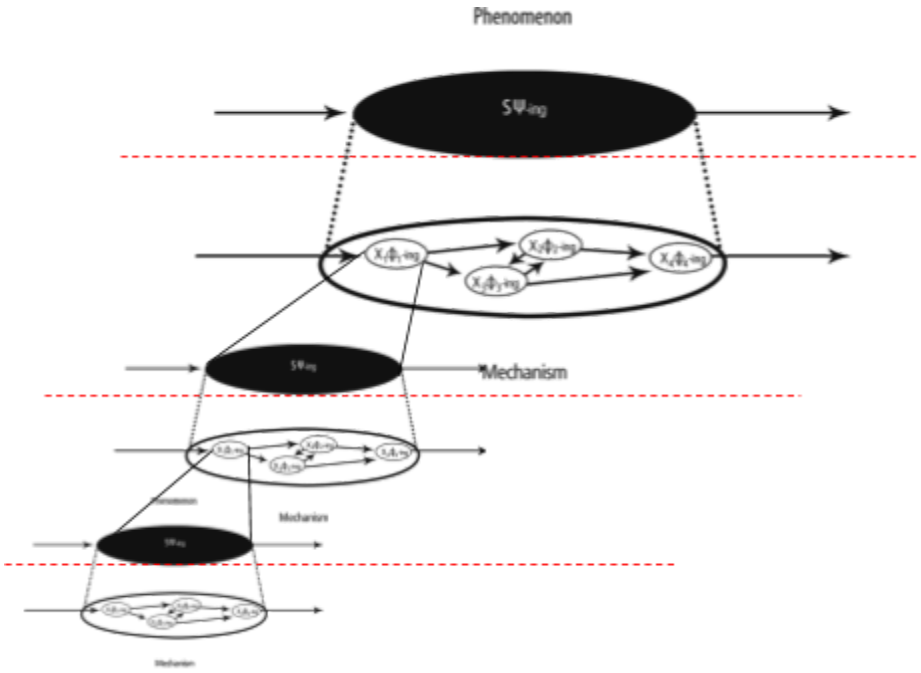


Figure 1 (Constitutive) Mechanistic Explanation (adapted from Craver 2007;7).

2. On this dominant form of mechanistic explanation - depicted in figure 1 - the phenomena ('S's ψ-ing') is explained by the temporal and spatial organisation of the *activities of the component entities*

[85]These two types of explanation are thus distinguished on the basis of the relationship - viz. causal or constitutive - obtaining between the explanandum and the explanans (Kaiser and Krickel (2016,7) and together reflect the two ways in which a phenomenon can be located within the causal nexus (Craver and Tabery 2017). Standard examples of etiological mechanistic explanations include neurotransmitter release (Craver 2007,22) - in which the release of neurotransmitters into the synaptic cleft is explained by its previous causes (Ca2+ channels opening, the influx of Ca2+ etc.) - along with explanations of adaptive evolutionary outcomes (Skipper and Milstein (2005); Barros (2008); Glennan (2009) and the formation of a gamete (Glennan (2002)). In contrast, examples of constitutive mechanistic explanations in neuroscience include the case of spatial memory discussed in Chapter 2 (Craver 2007) and the long-term potentiation of synapses of neurons Bechtel and Abrahamsen (2005).

(Xs' ϕ-ings) which make up the mechanism. These are interconnected via various causal relationships (arrows).

3. These acting component-entities are in turn explained by the identification and description of their underlying neural mechanism, such that mechanistic explanation gives rise to a hierarchy of *levels* or nested mechanisms (the red dotted lines), which are individuated locally on the basis of the individual components of a neural mechanism[86].
4. These components (Xs' ϕ-ings) of a given mechanism for S are determined and identified on the basis of two conditions (153-157):

   (i) spatiotemporal parthood; X must be a spatiotemporal part of the system whose behaviour is to be explained (Krickel 2017). And

   (ii) Xs' ϕ-ings and S's ψ-ing must be *mutually manipulable* viz. A part is constitutively relevant component in a mechanism if one can change the behavior of the mechanism as a whole by intervening to change the behavior of the component *and* one can change the behavior of the component by intervening to change the behavior of the mechanism as a whole[87].

5. Crucially, on this form of mechanistic explanation, the relationship between the explanandum (the higher level cognitive phenomenon) and the explanans (the neural mechanism) is said to be one of *constitution* (indicated by the black dotted lines in figure 1). That is, the acting entities which constitute the mechanism (Xs' ϕ-ings) are spatially and temporally *contained within* the phenomenon (S's ψ-ing) . This is indicated by the way in which phenomena are often said to be explained by their 'underlying mechanism' which 'exhibit' or are responsible for them.

The main claim which emerges as a result of a constitutive-mechanistic analysis of the NCC programme then, and that which I want to make clear going forward, is that this analysis implies that the NCC researchers - both in content and state-based approaches - are best understood as searching for the neural mechanisms responsible for consciousness in this *constitutive* non-causal sense (as described in 1-5). On a constitutive-mechanistic analysis of contemporary NCC research, the specific non-correlational relationship between the relevant

[86] That is, that "X's ' ϕ-ing is at a lower mechanistic level than S's ψ-ing if and only if X's ϕ-ing is a *component* of the mechanism for S's ψ-ing" (Craver 2007;189).
[87]Again, as explained previously, the viability of this criterion, which Craver characterises formally in terms of *Woodwardian interventions* (2007153-155) is hotly disputed (Leuridan 2012, Baumgartner & Gebharter 2016, Romero 2015). The main worry here is that there is a tension or inconsistency in Craver's account insofar as interventionism is an account of *causation* and thus not suited to an analysis of constitution relations. See Romero (2015) Baumgartner & Gebharter (2016) and Krickel (*forthcoming*) for ways of resolving this inconsistency and retaining the constitution relation along mutual manipulability lines.

(hierarchically ordered) neural mechanisms and the aspect of consciousness to be explained in each given case is one of constitution[88].

2. Motivating a Constitutive-Mechanistic Analysis of the NCC

As explained in Chapter 2, Neisser's argument in favour of his causal-mechanistic interpretation of the NCC programme was twofold: first, that such an analysis is plausible - and indeed is to be expected - from the perspective of philosophy of neuroscience, and second, that this analysis is supported by attention to the recent experimental paradigms employed in the search for content-specific NCCs. That is, Neisser argued that attention to the details of standard Binocular Rivalry paradigms supports his causal characterisation of the NCC programme insofar as the neuroscientists working in these paradigms are best understood as identifying the neural activations which *make the difference*, in the Woodwardian *causal* interventionist sense utilised by the new mechanists, to the experience of the preferred stimulus under examination (683,686). The distinction between etiological and constitutive forms of explanation, and reasons we might have for thinking that the etiological form mechanistic explanation is operational in this NCC context, are not discussed.

A closer look at the new mechanist literature however reveals that these two motivations that Neisser provides (in the absence of such etiological-compelling reasons), in fact provide stronger support for a *constitutive*-mechanistic analysis of the NCC research programme. Whilst etiological explanations are no doubt crucial in many areas of science (particularly evolutionary biology and related fields) including a small number of explanations in neuroscience, the *dominant* form of explanation which is discussed in this context - and that which Craver's influential (2007) treatment of neuroscientific explanation is primarily concerned with - is constitutive mechanistic explanation. In the (valid) interest of providing an account of what neuroscientists working on consciousness are up to which is congruent and supported by recent work in philosophy of neuroscience and neuroscientific explanation, we ought to first consider a constitutive mechanistic analysis of the NCC research. Second and more importantly, there are also good reasons for thinking that, on closer examination, the constitutive-mechanistic (as opposed to causal) analysis of the NCC is better reflected in the experimental paradigms just mentioned, insofar as the latter appear to exhibit central discovery strategies for identifying the component-acting- entities of *constitutive mechanisms* as detailed by the New Mechanists.

If a constitutive-mechanistic analysis of the NCC programme is correct, we should expect to see the broad explanatory mandate it recommends - viz. explanation via the identification of the neural mechanisms constitutive of consciousness (as outlined in 1-5) - reflected in the experimental practices and paradigms employed in recent empirical studies. In other words, these paradigms should reflect the search for constitutive

[88]I discuss this constitution relation in more detail in section 3.

mechanisms which underlie the various aspects of consciousness . As we saw previously, in his discussion, Neisser argued that researchers using Binocular Rivalry paradigms in the search for content NCCs - viz. those in which distinct stimuli are presented to the eyes of a conscious subject, causing the conscious experience to shift between the different stimuli every couple of seconds - demonstrate evidence of searching for the neural activations which make the difference, in Woodwardian interventionist terms (2002, Craver 2007) to the experience of the relevant stimulus under examination in the study[89]. Whilst Neisser argues that this provides evidence for a causal characterisation (and this prima facie would seem to do so) attention to Craver's account of constitutive mechanistic explanation -and in particular, his mutual manipulability criterion for constitutive relevance (4) which he formalises (not uncontroversially) in terms of *Woodwardian interventions* (153) - demonstrates that Neisser's claim is equally compatible with a constitutive analysis of the NCC programme.

Moreover, the common experimental paradigms used in NCC research just mentioned - namely, the report-based and no-report based Binocular Rivalry paradigms - in an NCC fit nicely into such a constitutive-mechanistic picture insofar as they appear to exemplify a top-down form of interlevel experiment (i.e. those used to describe constitutive mechanisms) which Craver (2007) and Kastner (2017;58) refer to as *'Activation Experiments'*. One of four types of experiments used to test for *constitutive relevance* relations among entities and higher level phenomena, the experimenter in an activation experiment seeks to manipulate S in order to elicit its ψ-ing - in this case, causing the conscious stimuli to shift - in order to observe and identify Xs' ϕ-ing viz. The brain activity recorded by fMRI in the fronto-parietal network and/or posterior cortical areas in these cases[90]. The constitutive-mechanistic analysis finds support in the experimental practices utilised in contemporary NCC practices.

3. Reductive Physicalism and Constitutive Mechanistic Explanation

Thus far I have outlined and motivated an alternative *constitutive-mechanistic* analysis of the NCC research programme. A full examination of the metaphysical consequences of this analysis for the metaphysics of mind and consciousness (that which would constitute the complete discipline-specific approach outlined in Chapter 3) goes well beyond the scope of this Chapter. Here in conclusion however, I want to briefly examine a consequence of the main claim concerning the relationship between the neural mechanism(s) and explanandum phenomenon that this constitutive-mechanistic analysis provides. As stated above, if this analysis of the NCC research project is broadly correct, it follows that the relationship between the relevant (hierarchically ordered)

[89] See Tong et al. (1998), Logothetis et al. (2002), Tsuchiya and Koch (2005), Breitmeyer and Ogmen (2000).
[90] Koch et al. (2016). Here for example, the recent move in the content-based NCC approach towards no-report based paradigms and efforts to 'screen off' irrelevant neural activity relating to selective attention, self-monitoring or report, that precede or follow content NCCs can also be understood in mechanistic terms viz. as the attempt to separate constitutive and non-constitutive components acting entities in the relevant constitutive mechanisms.

neural mechanism identified by a completed NCC programme and [the aspect(s) of] consciousness explained is one of *constitution.* Whilst there has been much recent discussion about the nature of this constitution relation and its connection to constitution as understood in traditional metaphysics, and related notions of realization[91], this mechanistic constitution relation is thought to *minimally* imply that the acting entities which constitute the mechanism (Xs' ϕ-ings) are spatially and temporally *contained within*, and give rise to, the phenomenon in question (S's ψ-ing)[92]. This is a substantial claim, and importantly, one which appears to be in tension with certain reductive type-identity varieties of physicalism currently debated in the metaphysics of mind which typically claim that mental state types are *identical* with brain states types [93]. The claim that I wish to put forward then, is that the mechanistic analysis of the NCC programme presented in this Chapter provides the foundation of a novel argument against such views, on the basis that such identity claims are precluded by the sort of metaphysical relationship commonly thought to hold between mechanisms and phenomena detailed by a mechanistic analysis of contemporary research in the neuroscience of consciousness.

Bibliography

[91]Harbecke (2010) and Gillett (2013).
[92] See Kaiser and Krickel (20168-9) for a defense of this minimal analysis which follows from Craver's account of constitutive relevance.
[93] See Chapter 1.1. Levin (1991), McLaughlin (1999), Bechtel and McCauley (1999), Polger (2004), Kim [conditionally] (1998, 2005). As Polger (2009;823) writes: "The identity theory holds that mental states are identical to brain states – not merely correlated with them, caused by them, realized by them, superadded to them, etc".

Adams, Robert Merrihew (2007). Idealism vindicated. In Peter van Inwagen & Dean Zimmerman (eds.), Persons: Human and Divine. Oxford University Press. pp. 35-54.

Alai, Mario (forthcoming). The Underdetermination of Theories and Scientific Realism. Axiomathes:1-17.

Alter, Torin & Nagasawa, Yujin (2012). What is Russellian Monism? Journal of Consciousness Studies 19 (9-10):9-10.

Aru, J., Bachmann, T., Singer, W. & Melloni, L. Distilling the neural correlates of consciousness. Neurosci. Biobehav. Rev. 36, 737–746 (2012).

Audi, Paul (2012). A clarification and defense of the notion of grounding. In Fabrice Correia & Benjamin Schnieder (eds.), Metaphysical Grounding: Understanding the Structure of Reality. Cambridge University Press. pp. 101-121.

Audi, Paul (2012). Grounding: Toward a Theory of the In-Virtue-Of Relation. Journal of Philosophy 109 (12):685-711.

Baars, Bernard J. (1997). In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. Journal of Consciousness Studies 4 (4):292-309.

Baars, Bernard J. (1999). Attention vs consciousness in the visual brain: Differences in conception, phenomenology, behavior, neuroanatomy, and physiology. Journal of General Psychology 126:224-33.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Prog. Brain Res.
Baars, B., Franklin, S., and Ramsoy, T. Z. (2012). Global workspace dynamics: cortical “binding and propagation” enables conscious contents. Front. Conscious. Res.

Balog, Katalin (1999). Conceivability, possibility, and the mind-body problem. Philosophical Review 108 (4):497-528.

Balog, Katalin (2012). In Defense of the Phenomenal Concept Strategy1. Philosophy and Phenomenological Research 84 (1):1-23.

Balog, Katalin, Physicalism, dualism, and metaphysical gridlock. Available online: http://andromeda.rutgers.edu/~kbalog/Web%20publications/ZRedux.pdf

Baumgartner, Michael ; Casini, Lorenzo & Krickel, Beate (2018). Horizontal Surgicality and Mechanistic Constitution. Erkenntnis:1-14.

Baumgartner, Michael & Gebharter, Alexander (2016). Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. British Journal for the Philosophy of Science 67 (3):731-756.

Bayne, Tim ; Hohwy, Jakob & Owen, Adrian M. (2016). Are There Levels of Consciousness? Trends in Cognitive Sciences 20 (6):405-413.

Bayne Tim; On the axiomatic foundations of the integrated information theory of consciousness, Neuroscience of Consciousness, Volume 2018, Issue 1, 1 January 2018, niy007, https://doi.org/10.1093/nc/niy007

Bechtel, William P. & McCauley, Robert N. (1999). Heuristic identity theory (or back to the future): The mind-body problem against the background of research strategies in cognitive neuroscience. In Martin Hahn & S. C. Stoness (eds.), Proceedings of the 21st Annual Meeting of the Cognitive Science Society. Lawrence Erlbaum. pp. 67-72.

Bechtel, William & Abrahamsen, Adele (2005). Explanation: a mechanist alternative. Studies in History and Philosophy of Biological and Biomedical Sciences 36 (2):421-441.

Bechtel, William (2005). Explanation: A mechanist alternative. Studies in History and Philosophy of Biol and Biomed Sci 36 (2):421--441.

Bechtel, William (2007). Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience. Psychology Press.

Beebee, Helen (2018). I—The Presidential AddressPhilosophical Scepticism and the Aims of Philosophy. Proceedings of the Aristotelian Society 118 (1):1-24.

Belot, Gordon (2013). Symmetry and Equivalence. In Robert Batterman (ed.), The Oxford Handbook of Philosophy of Physics. Oxford University Press. pp. 318-339.

Bennett, K. (2003). Why the exclusion problem seems intractable, and how, just maybe, to tract it. Noûs, 37(3), 471–497.

Bennett, Karen (2008). Exclusion again. In Jakob Hohwy & Jesper Kallestrup (eds.), Being Reduced: New Essays on Reduction, Explanation, and Causation. Oxford University Press. pp. 280--307.

Bernstein, Sara (2016). Grounding Is Not Causation. Philosophical Perspectives 30 (1):21-38.

Bird, Alexander (2007). Nature's metaphysics. Oxford University Press.

Block, Ned (2003). Do causal powers drain away. Philosophy and Phenomenological Research 67 (1):133-150

Block, Ned (2007). Consciousness, Accessibility, and the Mesh between Psychology and Neuroscience. Behavioral and Brain Sciences 30 (5):481--548.

Block, Ned (2007). Overflow, access, and attention. Behavioral and Brain Sciences 30 (5-6):530-548.

Block, Ned (2009). Comparing the major theories of consciousness. In Michael Gazzaniga (ed.), The Cognitive Neurosciences IV. pp. 1111-1123.

Block, Ned & Stalnaker, Robert (1999). Conceptual analysis, dualism, and the explanatory gap. Philosophical Review 108 (1):1-46.

Block, Ned ; Carmel, David ; Fleming, Stephen M. ; Kentridge, Robert W. ; Koch, Christof ; Lamme, Victor A. F. ; Lau, Hakwan & Rosenthal, David (2014). Consciousness science: real progress and lingering misconceptions.Trends in Cognitive Sciences 18 (11):556-557.

Boly, M. et al. Consciousness in humans and non-human animals: recent advances and future directions. Front. Psychol. 4, 625 (2013).

Boly, M. et al. Brain connectivity in disorders of consciousness. Brain Connect.2, 1–10 (2012).

Brown, E. N., Lydic, R. & Schiff, N. D. General anesthesia, sleep, and coma. N. Engl. J. Med. 363, 2638–2650 (2010).

Bruntrup, Godehard & Jaskolla, Ludwig (eds.) (2017). Panpsychism: Contemporary Perspectives. Oxford University Press USA.

Bryant, Amanda (2017). Keep the Chickens Cooped: The Epistemic Inadequacy of Free Range Metaphysics. Synthese:1-21.

Bryant, Amanda (2018). Naturalizing grounding: How theories of ground can engage science. Philosophy Compass:e12489.

Bucci, Alessio & Grasso, Matteo (2017). Sleep and dreaming in the predictive processing framework. Philosophy and Predictive Processing.

Butterfield, Jeremy (2014). On under-determination in cosmology. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 46 (1):57-69.

Carey, Brandon (2011). Overdetermination And The Exclusion Problem. Australasian Journal of Philosophy 89 (2):251-262.

Carruthers, Peter (2000). Phenomenal Consciousness: A Naturalistic Theory. Cambridge University Press.

Carruthers, Peter & Veillet, Benedicte (2007). The phenomenal concept strategy. Journal of Consciousness Studies 14 (s 9-10):212-236.

Cerullo, Michael (2011). Integrated Information Theory A Promising but Ultimately Incomplete Theory of Consciousness. Journal of Consciousness Studies 18 (11-12):11-12.

Cerullo MA (2015) The Problem with Phi: A Critique of Integrated Information Theory. PLoS Comput Biol 11(9):

Chakravartty, Anjan (2013). On the Prospects of Naturalized Metaphysics. In Don Ross, James Ladyman & Harold Kincaid (eds.), Scientific Metaphysics. Oxford University Press. pp. 27-50.

Changeux, Jean-Pierre & Dehaene, Stanislas (2005). Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattentional blindness. PLoS Biology 3 (5):e141.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. J. Consc. Stud. 2, 200–219.
Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, David J. (2000). What is a neural correlate of consciousness? In Thomas Metzinger (ed.), Neural Correlates of Consciousness. MIT Press. pp. 17--39.

Chalmers, David J. (2004). How can we construct a science of consciousness? In Michael S. Gazzaniga (ed.), The Cognitive Neurosciences Iii. MIT Press. pp. 1111--1119.

Chalmers, David John (2010). The Character of Consciousness. Oxford University Press.

Chalmers, David (forthcoming). Idealism and the Mind-Body Problem. In William Seager (ed.), The Routledge Handbook of Panpsychism. Routledge.

Chalmers, D. ; Manley, D. & Wasserman, R. (eds.) (2009). Metametaphysics: New Essays on the Foundations of Ontology. Oxford University Press.

Chalmers, David J. (2002). Does conceivability entail possibility? In Tamar S. Gendler & John Hawthorne (eds.), Conceivability and Possibility. Oxford University Press. pp. 145--200.

Chalmers, David, The Meta-Problem of Consciousness. Manuscript (forthcoming). Avaliable online: https://philpapers.org/archive/CHATMO-32.pdf

Chalmers, David J. (2007). Phenomenal concepts and the explanatory gap. In Torin Alter & Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford University Press.

Chalmers, David J. (2016). Referentialism and the Objects of Credence: A Reply to Braun. Mind 125 (498):499-510.

Chalmers, David J. & Jackson, Frank (2001). Conceptual analysis and reductive explanation. Philosophical Review 110 (3):315-61.

M. Chirimuuta; Explanation in Computational Neuroscience: Causal and Non-causal, The British Journal for the Philosophy of Science

Clark, Andy (2000). A case where access implies qualia? Analysis 60 (1):30-37.

Clark, A. (2013). Whatever next. Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. 36, 181–204. doi: 10.1017/S0140525X12000477

Clark, Andy (2016). Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press USA.

Clark et al. (2017) Expecting Ourselves: Embodied Prediction and the Construction of Conscious Experience (XSPECT).http://eidyn.ppls.ed.ac.uk/project/expecting-ourselves-embodied-prediction-and-construction-conscious-experience-xspect [accessed 07/17]

Conee, Earl (1994). Phenomenal knowledge. Australasian Journal of Philosophy 72 (2):136-150.

Crane, Tim et al. 'New Directions in the Study of Mind' [Information Booklet accessed 07/17] http://www.newdirectionsproject.com/about/

Crane, Tim & Mellor, D. H. (1990). There is No Question of Physicalism. Mind 99 (394):185-206.

Craver Carl F., (2007). Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience. Oxford University Press, Clarendon Press.

Craver, Carl F. (2005). Beyond reduction: mechanisms, multifield integration and the unity of neuroscience. Studies in History and Philosophy of Science Part C 36 (2):373-395.

Craver, Carl F. (2013). Functions and mechanisms: a perspectivalist view. In Philippe Huneman (ed.), Functions: Selection and Mechanisms. Springer. pp. 133--158.

Craver, Carl F. (2014). The Ontic Account of Scientific Explanation. In Marie I. Kaiser, Oliver R. Scholz, Daniel Plenge & Andreas Hüttemann (eds.), Explanation in the Special Sciences: The Case of Biology and History. Springer Verlag. pp. 27-52.

Craver, Carl and Tabery, James, "Mechanisms in Science", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/>.

Crick, Francis & Koch, Christof (1990). Toward a neurobiological theory of consciousness. Seminars in the Neurosciences 2:263-275.

Crick, Francis & Koch, Christof (1998). Consciousness and neuroscience. Cerebral Cortex.

Crick, Francis & Koch, Christof (2007). A neurobiological framework for consciousness. In Max Velmans & Susan Schneider (eds.), The Blackwell Companion to Consciousness. Blackwell. pp. 567--579.

Crick, Francis (1994). The Astonishing Hypothesis: The Scientific Search for the Soul. Scribners.

Crick, F., and Koch, C. (2003). A framework for consciousness. Nat. Neurosci. 6, 119–126. doi: 10.1038/nn0203-119

Crisp, Thomas M. & Warfield, Ted A. (2001). Jaegwon Kim, Mind in a Physical World. Noûs 35 (2):304-316.

Dasgupta, Shamik (2014). The Possibility of Physicalism. Journal of Philosophy 111 (9-10):557-592.

Davidson, Donald (1970). Mental Events. In L. Foster & J. W. Swanson (eds.), Essays on Actions and Events. Clarendon Press. pp. 207-224.

De Brigard, Felipe (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. Synthese 191 (2):1-31.

De Brigard, F. (2014). Self-Stultification Objection. Journal of Consciousness Studies 21 (5-6):120-130.

Dehaene, Stanislas & Changeux, Jean-Pierre (2004). Neural Mechanisms for Access to Consciousness. In Michael S. Gazzaniga (ed.), The Cognitive Neurosciences. MIT Press. pp. 1145-1157.

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79, 1–37.

Dehaene, S. & Changeux, J.-P. Experimental and theoretical approaches to conscious processing. Neuron 70, 200–227 (2011).

Dennett, Daniel C. (1991). Consciousness Explained. Penguin Books.

Dennett, Daniel C. (2001) The fantasy of first-person science. https://ase.tufts.edu/cogstud/dennett/papers/chalmersdeb3dft.htm [accessed 07/17]

Diaz-Leon, E. (2008). Defending the phenomenal concept strategy. Australasian Journal of Philosophy 86 (4):597 – 610.

Diaz-Leon, E. (2010). Can Phenomenal Concepts Explain The Epistemic Gap? Mind 119 (476):933-951.

Dienes Z. Subjective measures of unconscious knowledge. Prog. Brain Res. 2008; 168: 49-64

Dretske, Fred (1995). Naturalizing the Mind. MIT Press.

Edelman, G. M. (1989). The Remembered Present: A Biological Theory of Consciousness. New York, NY: Basic Books.

Ellis, Brian (2001). Scientific Essentialism. Mind 113 (450):334-340.

Endicott, Ronald P. (2012). Resolving arguments by different conceptual traditions of realization. Philosophical Studies 159 (1):41-59.

Eronen, M.I., 2013, “No Levels, No Problems: Downward Causation in Neuroscience”, Philosophy of Science, 80: 1042–1052.

Eronen, M.I., 2015, “ Levels of Organization: A Deflationary Account”, Biology and Philosophy, 30: 39–58

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2009). Phenomenological architecture of a mind and operational architectonics of the brain: the unified metastable continuum. New Math. Nat. Comput. 5, 221–244.

Fingelkurts, A. A., Fingelkurts, A. A., and Neves, C. F. H. (2010). Natural world physical, brain operational, and mind phenomenal space-time. Phys. Life Rev. 7, 195–249.

Fink, Sascha Benjamin (2016). A Deeper Look at the "Neural Correlate of Consciousness". Frontiers in Psychology 7.

Frankish, Keith (2012). Quining diet qualia. Consciousness and Cognition 21 (2):667-676.

Frankish, K. (2016). “Illusionism as a Theory of Consciousness, Journal of Consciousness Studies 23(11-12).

Fodor, Jerry (1974). Special Sciences, or Disunity of Science as a Working Hypothesis. Synthese 28 (2):97--115.

Fodor, Jerry (1974). Special Sciences, or Disunity of Science as a Working Hypothesis. Synthese 28 (2):97--115.

French, Steven (2014). The Structure of the World: Metaphysics and Representation. Oxford University Press.

French, Steven (2018). Toying with the Toolbox: How Metaphysics Can Still Make a Contribution. Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie 49 (2):211-230.

French, Steven & Ladyman, James (1999). Reinflating the semantic approach. International Studies in the Philosophy of Science 13 (2):103 – 121.

Gatzia, Dimitria Electra & Brogaard, Brit (2016). What Can Neuroscience Tell Us about 64 65 the Hard Problem of Consciousness? Frontiers in Neuroscience 10:395.

Gennaro, Rocco J. (1996). Consciousness and Self-Consciousness: A Defense of the Higher-Order Thought Theory of Consciousness. John Benjamins.

Gibb, Sophie (2015). The Causal Closure Principle. Philosophical Quarterly 65 (261):626-647.

Gillett, Carl (2002). The metaphysics of realization, multiple realizability, and the special sciences. Journal of Philosophy 100 (11):591-603.

Glennan, StuartS (1996). Mechanisms and the nature of causation. Erkenntnis 44 (1):49--71.

Glennan, Stuart (2005). Modeling mechanisms. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 36 (2):443-464.

Glennan, Stuart (2017). The New Mechanical Philosophy. Oxford University Press.

Glennan, Stuart (2005). Modeling mechanisms. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 36 (2):443-464.

Glennan, Stuart (2009). Mechanisms. In Helen Beebee, Christopher Hitchcock & Peter Menzies (eds.), The Oxford Handbook of Causation. Oxford University Press.

Godfrey-Smith, Peter (2006). The strategy of model-based science. Biology and Philosophy 21 (5):725-740.

Godfrey-Smith, Peter (2006). Theories and Models in Metaphysics. The Harvard Review of Philosophy14 (1):4-19.
Goff, Philip (2009). Why Panpsychism doesn't Help Us Explain Consciousness. Dialectica 63 (3):289-311.

Goff, Philip (2011). A posteriori physicalists get our phenomenal concepts wrong. Australasian Journal of Philosophy 89 (2):191 - 209.

Goff, Philip (2017). Consciousness and Fundamental Reality. Oxford University Press.

Goff, Philip and Coleman, Sam. 'Russellian Monism' (forthcoming). In U. Kriegel (ed.), Oxford Handbook of the Philosophy of Consciousness. Oxford University Press.

Gosseries, O., Di, H., Laureys, S. & Boly, M. Measuring consciousness in severely damaged brains. Annu. Rev. Neurosci. 37, 457–478 (2014).

Guay, Alexandre & Pradeu, Thomas (forthcoming). Right out of the box: How to situate metaphysics of science in relation to other metaphysical approaches. Synthese:1-20.

Hall, Ned (2004). Two concepts of causation. In John Collins, Ned Hall & Laurie Paul (eds.), Causation and Counterfactuals. MIT Press. pp. 225-276.

Harbecke, Jens (2010). Mechanistic Constitution in Neurobiological Explanations. International Studies in the Philosophy of Science 24 (3):267-285.

Harbecke, Jens (2014). The role of supervenience and constitution in neuroscientific research. Synthese 191 (5):1-19.

Havlík, Marek ; Kozáková, Eva & Horáček, Jiří (2017). Why and How. The Future of the Central Questions of Consciousness. Frontiers in Psychology 8.

Hempel, Carl G. & Oppenheim, Paul (1948). Studies in the logic of explanation. Philosophy of Science 15 (2):135-175.

Hempel, Carl (1965). Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. The Free Press.

Hill, Christopher S. & Mclaughlin, Brian P. (1999). There are fewer things in reality than are dreamt of in Chalmers's philosophy. Philosophy and Phenomenological Research 59 (2):445-454.

Hohwy, J. (2007). The search for neural correlates of consciousness. Philosophy Compass, 2(3), 461–474.

Hohwy, Jakob (2009). The neural correlates of consciousness: New experimental approaches needed? Consciousness and Cognition 18 (2):428-438.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. Front. Psychol. 3:96. doi: 10.3389/fpsyg.2012.00096

Hohwy, Jakob (2013). The Predictive Mind. Oxford University Press UK.

Hohwy, Jakob & Bayne, Timothy (2015) The neural correlates of consciousness: Causes, confounds and constituents.

Horgan, Terence E. (1984). Jackson on physical information and qualia. Philosophical Quarterly 34 (April):147-83.

Horgan, Terence (1984). Functionalism, qualia, and the inverted spectrum. Philosophy and Phenomenological Research 44 (June):453-69.

Horgan, Terence E. (1997). Kim on mental causation and causal exclusion. Philosophical Perspectives 11 (s11):165-84.

Horgan, Terence E. (2001). Causal compatibilism and the exclusion problem. Theoria 16 (40):95-116.

Howell, Robert J. (2013). Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism. Oxford University Press.

Humphrey, Nicholas (2011). Soul Dust: The Magic of Consciousness. Princeton University Press.
Illari, Phyllis (2013). Mechanistic Explanation: Integrating the Ontic and Epistemic. Erkenntnis 78 (2):237-255.

Illari, Phyllis & Williamson, Jon (2012). What is a mechanism? Thinking about mechanisms across the sciences. European Journal for Philosophy of Science 2 (1):119-135.

Jackson, Frank (1998). From Metaphysics to Ethics: A Defence of Conceptual Analysis. Oxford University Press.

Jackson, Frank (2007). A priori physicalism. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), Contemporary Debates in Philosophy of Mind. Blackwell.

Kaiser, M. and C.F. Craver, 2013, "Mechanisms and Laws: Clarifying the Debate", in Hsiang-Ke Chao, Szu-Ting Chen and Roberta L. Millstein (eds), Mechanism and Causality in Biology and Economics, Dordrecth: Springer, pp. 125–145.

Kaiser, Marie I. & Krickel, Beate (2016). The Metaphysics of Constitutive Mechanistic Phenomena. British Journal for the Philosophy of Science:axv05

Kallestrup, Jesper (2006). The causal exclusion argument. Philosophical Studies 131 (2):459-85.

Kammerer, François (2018). Can you believe it? Illusionism and the illusion meta-problem. Philosophical Psychology 31 (1):44-67.

Kästner, Lena & Andersen, Lise Marie (forthcoming). Intervening into mechanisms: Prospects and challenges. Philosophy Compass.

Kästner, Lena, Philosophy of Cognitive Neuroscience: Causal Explanations, Mechanisms and Experimental Manipulations.

Kastrup, Bernardo (2017). On the Plausibility of Idealism: Refuting Criticisms. Disputatio 9 (44):13-34.

Kim, Jaegwon (1993). Supervenience and Mind: Selected Philosophical Essays. Cambridge University Press.

Kim, Jaegwon (1998). Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation. MIT Press.

Kim, Jaegwon (2005). Physicalism, or Something Near Enough. Princeton University Press.

Kim, Jaegwon (2007). Causation and mental causation. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), Contemporary Debates in Philosophy of Mind. Blackwell. pp. 227--242.

Kirk, Robert & Squires, J. E. R. (1974). Zombies v. Materialists. Aristotelian Society Supplementary Volume 48 (1):135-164.

Kitcher, Philip (1989). Explanatory unification and the causal structure of the world. In Philip Kitcher & Wesley Salmon (eds.), Scientific Explanation. Minneapolis: University of Minnesota Press. pp. 410-505.

Koch, Christof (2004). The Quest for Consciousness. Roberts & Company.

Koch, Christof (2012). Consciousness: Confessions of a Romantic Reductionist. MIT Press.

Koch, Christof et. al (2016). Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience 17,307–321

Koch, C., and Tsuchiya, N. (2007). Attention and consciousness: two distinct brain processes. Trends Cogn. Sci. 11, 16–22.

Koslicki, Kathrin (2016). Where grounding and causation part ways: comments on Schaffer. Philosophical Studies 173 (1):101-112.

Kovacs, David Mark (forthcoming). Grounding and the argument from explanatoriness. Philosophical Studies:1-26.

Kriegel, Uriah (forthcoming). Beyond the Neural Correlates of Consciousness. In U. Kriegel (ed.), Oxford Handbook of the Philosophy of Consciousness. Oxford University Press.

Krickel, B., 2014, The Metaphysics of Mechanism, PhD Dissertation. Humboldt-Universität zu Berlin.
Krickel, Beate (2018). Saving the mutual manipulability account of constitutive relevance. Studies in History and Philosophy of Science Part A 68:58-67.

Krickel, Beate (2017). Making Sense of Interlevel Causation in Mechanisms from a Metaphysical Perspective. Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie 48 (3):453-468.

Krickel, Beate (2017). A Regularist Approach to Mechanistic Type-Level Explanation. British Journal for the Philosophy of Science:00-00.

Kripke, Saul (1980). Naming and Necessity. Harvard University Press.

Kroedel, Thomas & Schulz, Moritz (2016). Grounding mental causation. Synthese 193 (6):1909-1923.

Kroedel, Thomas (2015). Dualist Mental Causation and the Exclusion Problem. Noûs 49 (2):357-375.

Ladyman, James (2012). Science, metaphysics and method. Philosophical Studies 160 (1):31-51.

Ladyman, James & Ross, Don (2007). Every Thing Must Go: Metaphysics Naturalized. Oxford University Press.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. Trends Cogn. Sci. 10, 494–501.

Lamme, V. A. (2010). How neuroscience will change our view on consciousness. Cogn. Neurosci. 1, 204–220.
Lau, Hakwan & Rosenthal, David (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences 15 (8):365-373.

Laudan, Larry & Leplin, Jarrett (1991). Empirical equivalence and underdetermination. Journal of Philosophy 88 (9):449-472.

Laureys, S., Owen, A. M. & Schiff, N. D. Brain function in coma, vegetative state, and related disorders.Lancet Neurol. 3, 537–546 (2004)

Laureys, S. (2005). The neural correlate of (un)awareness: Lessons from the vegetative state. Trends in Cognitive Sciences, 9(12), 556–559.
Laureys, S., and Boly, M. (2008). The changing spectrum of coma. Nat. Clin. Pract. Neurol. 544–546.
Leuridan, Bert (2012). Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms. British Journal for the Philosophy of Science 63 (2):399-427.

Levin, Janet (1991). Analytic functionalism and the reduction of phenomenal states. Philosophical Studies 61 (March):211-38.

Levine, Joseph (1983). Materialism and qualia: The explanatory gap. Pacific Philosophical Quarterly 64 (October):354-61.

Levine, Joseph (2001). Phenomenal consciousness and the first-person. PSYCHE: An Interdisciplinary Journal of Research On Consciousness 7.

Levine, Joseph (2001). Purple Haze: The Puzzle of Consciousness. Oxford University Press USA.

Levine, Joseph (2010). Demonstrative thought. Mind and Language 25 (2):169-195.

Lewis, David K. (1983). Philosophical Papers. Oxford University Press.

Lewis, David K. (1966). An Argument for the Identity Theory. Journal of Philosophy 63 (1):17-25.

List, Christian & Menzies, Peter (2017). My brain made me do it: The exclusion argument against free will, and what's wrong with it. In H. Beebee, C. Hitchcock & H. Price (eds.), Making a Difference. Oxford: Oxford University Press.

Loar, Brian (1990). Phenomenal states. Philosophical Perspectives 4:81-108.

Loar, Brian (2003). Phenomenal intentionality as the basis of mental content. In Martin Hahn & B. Ramberg (eds.), Reflections and Replies: Essays on the Philosophy of Tyler Burge. MIT Press. pp. 229--258.

Loewer, Barry M. (1995). An argument for strong supervenience. In Elias E. Savellos & U. Yalcin (eds.), Supervenience: New Essays. Cambridge University Press. pp. 218--225.

Loewer, Barry M. (2007). Mental causation, or something near enough. In Brian P. McLaughlin & Jonathan D.

Lowe, E. J. (2006). Subjects of Experience. Cambridge University Press.

Lowe, E. J. (2006). Non-cartesian substance dualism and the problem of mental causation. Erkenntnis 65 (1):5-23.

Lowe, E. J. (2011). The rationality of metaphysics. Synthese 178 (1):99-109.

Lycan, William G. (1996). Consciousness and Experience. MIT Press

Machamer, Peter (2004). Activities and causation: The metaphysics and epistemology of mechanisms. International Studies in the Philosophy of Science 18 (1):27 – 39.

Machamer, Peter K. ; Darden, Lindley & Craver, Carl F. (2000). Thinking about mechanisms. Philosophy of Science 67 (1):1-25.

Mackie, J. L. (1974). The Cement of the Universe. Oxford, Clarendon Press.

Maddy, Penelope (2007). Second Philosophy: A Naturalistic Method. Oxford University Press.

Massimini, M., Boly, M., Casali, A., Rosanova, M., and Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. Prog. Brain Res. 177, 201–214

Massimini, M. et al. Breakdown of cortical effective connectivity during sleep. Science 309, 2228–2232 (2005).
Massimini, M., Ferrarelli, F., Esser, S. K., Riedner, B. A., Huber, R., Murphy, M., et al. (2007). Triggering sleep slow waves by transcranial magnetic stimulation. Proc. Natl. Acad

Massimini, M. et al. Cortical reactivity and effective connectivity during REM sleep in humans. Cogn. Neurosci. 1, 176–183 (2010).

Massimini, M and Tononi, G. 'Sizing up Consciousness: Towards an Objective Measure of the Capacity for Conscious Experience '.(2018). Oxford Univeristy Press.

Maudlin, Tim (2007). The Metaphysics Within Physics. Oxford University Press.

McLaughlin, Brian P. (2016). Hill on phenomenal consciousness. Philosophical Studies 173 (3):851-860.

Melnyk, Andrew (1997). How to keep the 'physical' in physicalism. Journal of Philosophy 94 (12):622-637.

Melnyk, Andrew (2003). A Physicalist Manifesto: Thoroughly Modern Materialism. Cambridge University Press.

Mendelovici, Angela (forthcoming). Panpsychism's Combination Problem Is a Problem for Everyone. In William Seager (ed.), The Routledge Handbook of Panpsychism. London, UK: Routledge.

Metzinger, Thomas (ed.) (2000). Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press.

Michel, M. ASSC 22 Tutorial on History and Philosophy of Science Perspectives on Consciousness. 'Consciousness Science: A History of Unsolved Debates'(unpublished).

Miller, S. M. Closing in on the constitution of consciousness. Front. Psychol. 5, 1293 (2014)

Mindt, Garrett (2017). The Problem with the 'Information' in Integrated Information Theory. Journal of Consciousness Studies 24 (7-8):130-154.

Montero, Barbara (2001). Post-physicalism. Journal of Consciousness Studies 8 (2):61-80.

Montero, Barbara Gail (2012). Irreverent Physicalism. Philosophical Topics 40 (2):91-102.

Montero, Barbara (1999). The body problem. Noûs 33 (2):183-200.

Montero, Barbara (1999). The Body of the Mind-Body Problem. Annals of the Japan Association for Philosophy of Science 9 (4):207-217.

Montero, Barbara & Papineau, David (2005). A defense of the via negativa argument for physicalism. Analysis 65 (3):233-237.

Moore, Dwayne (2014). The Epistemic Argument for Mental Causation. Philosophical Forum 45 (2):149-168.

Mørch, Hedda Hassel (2018). Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? Erkenntnis:1-21.

Morganti, Matteo & Tahko, Tuomas E. (2017). Moderately Naturalistic Metaphysics. Synthese 194 (7):2557-2580.

Mumford, Stephen & Tugby, Matthew (eds.) (2013). Metaphysics and Science. Oxford University Press.

Nagel, Thomas (1974). What is it like to be a bat? Philosophical Review 83 (October):435-50.

Nagel, Thomas (2012). Mind and Cosmos. Oxford University Press.

Neisser, Joseph (2012). Neural correlates of consciousness reconsidered. Consciousness and Cognition 21 (2):681-690.

Nemirow, Laurence (1990). Physicalism and the cognitive role of acquaintance. In William G. Lycan (ed.), Mind and Cognition. Blackwell.

Ney, Alyssa (2012). Neo-positivist metaphysics. Philosophical Studies 160 (1):53-78.

Ney, Alyssa (2008). Physicalism as an attitude. Philosophical Studies 138 (1):1 - 15.

Ney, Alyssa (2008). Defining physicalism. Philosophy Compass 3 (5):1033-1048.

Nida-Rümelin, Martine (2007). Transparency of experience and the perceptual model of phenomenal awareness. Philosophical Perspectives 21 (1):429–455.

Nida-Rümelin, Martine (2007). Dualist emergentism. In Brian P. McLaughlin & Jonathan D. Cohen (eds.), Contemporary Debates in Philosophy of Mind. Blackwell.

Noë, Alva & Thompson, Evan (2004a). Are there neural correlates of consciousness? Journal of Consciousness Studies 11 (1):3-28.

Noe, Alva & Thompson, Evan (2004b). Sorting out the neural basis of consciousness: Authors' reply to commentators. Journal of Consciousness Studies 11 (1):87-98.

Noë, A. & Thompson, E. (forthcoming). Neural correlates of consciousness and the matching-content doctrine. Journal of Consciousness Studies.

Nolan, Daniel (2015). The A Posteriori Armchair. Australasian Journal of Philosophy 93 (2):211-231.

Oizumi, M., Albantakis, L. & Tononi, G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. 10, e1003588 (2014).
Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., and Pickard, J. D. (2006). Detecting awareness in the vegetative state. Science 313, 1402

Owen, A. (2008). Disorders of consciousness. Annals of the New York Academy of Sciences, 1124, 225–238.

Owen, A. M. (2013). Detecting consciousness: a unique role for neuroimaging. Annu. Rev. Psychol. 64, 109–133.

Papineau, David (1993). Physicalism, consciousness, and the antipathetic fallacy. Australasian Journal of Philosophy 71 (2):169-83.

Papineau, David (1993). Philosophical Naturalism. Blackwell.

Papineau, David (1998). Mind the Gap. Noûs 32 (S12):373-388.

Papineau, David (2000). 10 The rise of physicalism. In M. W. F. Stone & Jonathan Wolff (eds.), The Proper Ambition of Science. Routledge. pp. 2--174.

Papineau, David (2001). The rise of physicalism. In Carl Gillett & Barry M. Loewer (eds.), Physicalism and its Discontents. Cambridge University Press.

Papineau, David (2002). Thinking About Consciousness. Oxford University Press UK.

Papineau, David (2008). Explanatory gaps and dualist intuitions. In Lawrence Weiskrantz & Martin Davies (eds.), Frontiers of Consciousness. Oxford University Press. pp. 2008--55.

Papineau, David, "Naturalism", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL <https://plato.stanford.edu/archives/win2016/entries/naturalism/>.

Papineau, David 'The Problem of Consciousness'. Forthcoming, In U. Kriegel (ed.), Oxford Handbook of the Philosophy of Consciousness. Oxford University Press.

Paul, L. A. (2012). Metaphysics as modeling: the handmaiden's tale. Philosophical Studies 160 (1):1-29.

Pautz, Adam. Is physicalism simpler than dualism? (draft).

Pautz, Adam, Can Russellian Monism Solve the Mind-Body Problem? (draft).

Pereboom, Derk (2002). Robust nonreductive materialism. Journal of Philosophy 99 (10):499-531.

Place, Ullin T. (1956). Is consciousness a brain process. British Journal of Psychology 47 (1):44-50.

Polger, Thomas (2004). Natural Minds. Bradford.

Polger, Thomas W. (2009). Evaluating the evidence for multiple realization. Synthese 167 (3):457 - 472.

Polger, Thomas W. (2011). Are sensations still brain processes? Philosophical Psychology 24 (1):1-21.

Polger, Thomas W. (2009). Identity theories. Philosophy Compass 4 (5):822-834.

Polger, Thomas W. & Shapiro, Lawrence A. (2016). The Multiple Realization Book. Oxford University Press UK.

Psillos, Stathis (1999). Scientific Realism: How Science Tracks Truth. Routledge.

Psillos, Stathos, Causal Pluralism (2009) available online:
.http://users.uoa.gr/%7Epsillos/PapersI/26-Causal%20Pluralism.pdf

Putnam, Hilary (1967). The nature of mental states. In W.H. Capitan & D.D. Merrill (eds.), Art, Mind, and Religion. Pittsburgh University Press. pp. 1--223.

Putnam, Hilary (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (eds.), Art, Mind, and Religion. University of Pittsburgh Press. pp. 37--48.

Putnam, H. (1973). Reductionism and the nature of psychology. Cognition, 2, 131–146.

Putnam, Hilary (1975). Philosophy and our mental life. In Mind, Language, and Reality. Cambridge University Press.

Raatikainen, Panu, Mental causation, interventions, and contrasts (2006).

Reutlinger, Alexander & Saatsi, Juha (2017). Introduction: Scientific Explanation Beyond Causation. In Alexander

Reutlinger & Juha Saatsi (eds.), Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations. Oxford: Oxford University Press.

Revonsuo, Antti (2000). Inner Presence: Consciousness As a Biological Phenomenon. MIT Press.

Revonsuo, Antti (2000). Prospects for a scientific research program on consciousness. In Thomas Metzinger (ed.), Neural Correlates of Consciousness. MIT Press.

Robinson, Howard (1982). Matter and Sense: A Critique of Contemporary Materialism. Cambridge University Press.

Robinson, William S. (1982). Causation, sensations, and knowledge. Mind 91 (October):524-40.

Robinson, Howard M. (ed.) (1993). Objections to Physicalism. Oxford University Press.

Robinson, William S. (2006). Knowing epiphenomena. Journal of Consciousness Studies 13 (1-2):85-100.

Robinson, William S. (2006). Understanding Phenomenal Consciousness. Philosophical Quarterly 56 (222):142-144.

Robinson, William S. (2018). Russellian Monism and Epiphenomenalism. Pacific Philosophical Quarterly 99 (1):100-117.

Roche, Michael (2014). Causal Overdetermination and Kim's Exclusion Argument. Philosophia 42 (3):809-826.

Romero, Felipe (2015). Why there isn't inter-level causation in mechanisms. Synthese 192 (11):3731-3755.

Rosen, Gideon (2010). Metaphysical Dependence: Grounding and Reduction. In Bob Hale & Aviv Hoffmann (eds.), Modality: Metaphysics, Logic, and Epistemology. Oxford University Press. pp. 109-36.

Rosenthal, David M. (1997). A theory of consciousness. In Ned Block, Owen J. Flanagan & Guven Guzeldere (eds.), The Nature of Consciousness. MIT Press.

Ross, Don ; Ladyman, James & Kincaid, Harold (eds.) (2013). Scientific Metaphysics. Oxford University Press.

Ryle, Gilbert (1949). The Concept of Mind. Hutchinson & Co.

Schaffer, Jonathan (2003). Overdetermining causes. Philosophical Studies 114 (1-2):23 - 45.

Searle, J. R. (2005). Consciousness: What we still don't know. The New York Review of Books, 52(1).

Seth, A.K. (2009) Explanatory correlates of consciousness: Theoretical and computational challenges. Cognitive Computing, 1, 50-63. http://dx.doi.org/10.1007/s12559-009-9007-x

Seth AK. The Grand Challenge of Consciousness. Frontiers in Psychology. 2010;1:5. doi:10.3389/fpsyg.2010.00005.

Seth, A. K., and Critchley, H. D. (2013). Extending predictive processing to the body: emotion as interoceptive inference. Behav. Brain Sci. 36, 227–228

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. Trends Cogn. Sci. 12, 314–321

Seth, A. K., Barrett, A. B. & Barnett, L. Causal density and integrated information as measures of conscious level. Philos. Trans. A Math. Phys. Eng. Sci. 369, 3748–3767 (2011).

Shanahan, M. (2008). A spiking neuron model of cortical broadcast and competition. Consciousness and Cognition 17 (1):288-303.

Shapiro, L. & Sober, E. (2012). Against proportionality. Analysis 72 (1):89-93.

Shapiro, Lawrence A. & Sober, Elliott (forthcoming). Epiphenomenalism - the do's and the don 'ts'. In G. Wolters & Peter K. Machamer (eds.), Studies in Causality: Historical and Contemporary. University of Pittsburgh Press.

Sheredos, Benjamin (2016). Re-reconciling the Epistemic and Ontic Views of Explanation. Erkenntnis 81 (5):919-949.

Siclari, F., LaRocque, J. J., Bernardi, G., Postle, B. R. & Tononi, G. The neural correlates of consciousness in sleep: a no-task, within-state paradigm. Preprint athttp://biorxiv.org/content/early/2014/12/30/012443(2014).

Sider, Theodore (2003). Review: What's so Bad about Overdetermination? Philosophy and Phenomenological Research 67 (3):719 - 726.

Smart, Jjc (1959). Sensations and brain processes. Philosophical Review 68 (April):141-56.

Soom, Patrice (2012). Mechanisms, determination and the metaphysics of neuroscience. Studies in History and Philosophy of Science Part C 43 (3):655-664.

Soto, Cristian (2017). Taking stock of the metaphysics of science debate: drawing disciplinary frontiers. Metascience 26 (2):1-5.

Soto, Cristian (2015). The current state of the metaphysics of science debate. Philosophica 90.

Spurrett, David & Papineau, David (1999). A note on the completeness of "physics". Analysis 59 (1):25-29.

Stanford, P. Kyle (2006). Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford University Press.

Stoljar, Daniel (2001). Physicalism. Stanford Encyclopedia of Philosophy.

Stoljar, Daniel (2010). Physicalism. Routledge.

Stoljar, Daniel & List, Christian (2017). Does the exclusion argument put any pressure on dualism? Australasian Journal of Philosophy 95 (1):96-108.

Strawson, Galen (2006). Realistic monism - why physicalism entails panpsychism. Journal of Consciousness Studies 13 (10-11):3-31.

Strawson, Galen (2016). Mind and Being: The Primacy of Panpsychism. In Godehard Brüntrup & Ludwig Jaskolla (eds.), Panpsychism: Contemporary Perspectives. Oxford University Press. pp. 000-00.

Suppes, Patrick (2002). Representation and Invariance of Scientific Structures.

Tahko, Tuomas E. (2015). An Introduction to Metametaphysics. Cambridge University Press.

Thomson-Jones, Martin (1997). Models and the Semantic View. Philosophy of Science 73 (5):524-535

Tiehen, Justin (forthcoming). Recent Work on Physicalism. Analysis.

Tiehen, Justin (2016). Physicalism Requires Functionalism: A New Formulation and Defense of the Via Negativa. Philosophy and Phenomenological Research 93 (1):3-24.

Tong, F. 'Competing Theories of Binocular Rivalry: A Possible Resolution'. Brain Mind 2 (2001):55–83.
——, et al. 'Binocular Rivalry and Visual Awareness in Human Extrastriate Cortex'. Neuron 21
1998): 753–9.

Tong, F., & Engel, S. A. (2001). Interocular rivalry revealed in the human cortical blind-spot representation. Nature, 411, 195–199.

Tong, F., Meng, M., and Blake, R. (2006). Neural bases of binocular rivalry. Trends Cogn. Sci. 10, 502–511

Tononi, G. An information integration theory of consciousness. BMC Neurosci. 5, 42 (2004).

Tononi, G. & Koch, C. Consciousness: here, there, and everywhere? Phil. Trans. R. Soc. B http://dx.doi.org/10.1098/rstb.2014.0167 (2015).

Tononi, Giulio Srinivasan (2007). The information integration theory of consciousness. In Max Velmans & Susan Schneider (eds.), The Blackwell Companion to Consciousness. Blackwell.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. Biol. Bull. 215, 216–242.

Tononi, G., and Koch, C. (2008). The neural correlates of consciousness: an update. Ann. N.Y. Acad. Sci. 1124, 239–261.

Tononi, G., and Massimini, M. (2008). Why does consciousness fade in early sleep. Ann. N.Y. Acad. Sci. 1129, 330–334.

Tononi, G. The integrated information theory of consciousness: an updated account. Arch. Ital. Biol. 150, 56–90 (2012).

Tononi, G. Integrated information theory. Scholarpediahttp://dx.doi.org/10.4249/scholarpedia.4164 (2015).

Tononi, G, Boly, M, Massimini, M, Koch, C. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience volume 17, pages 450–461 (2016).http://www.nature.com/articles/nrn.2016.

Tulodziecki, Dana (2013). Underdetermination, methodological practices, and realism. Synthese 190 (17):3731-3750.

Tulodziecki, Dana (2012). Epistemic Equivalence and Epistemic Incapacitation. British Journal for the Philosophy of Science 63 (2):313-328.

Tsuchiya, N., Wilke, M., Frässle, S. & Lamme, V. A. No-report paradigms: extracting the true neural correlates of consciousness. Trends Cogn. Sci. 19, 757–770 (2015).

Tye, Michael (1995). Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. MIT Press.

Tsuchiya, N., and Koch, C. (2005). Continuous flash suppression reduces negative afterimages. Nat. Neurosci. 8, 1096–1101.

van Fraassen, Bas C. (2002). The Empirical Stance. Yale University Press.

van Fraassen, Bas (2002). Science as representation: Flouting the criteria. Philosophy of Science 71 (5):794-804.

van Fraassen, Bas C. (1980). Formal Semantics and Logic. Journal of Symbolic Logic 45 (2):376-377.

van Fraassen, Bas C. (1980). Theory Construction and Experiment: An Empiricist View. PSA:Proceedings of the Biennial Meeting of the Philosophy of Science Association 1980:663-678.

Vernazzani, Alfredo (2015). Manipulating the Contents of Consciousness. Proceedings of the 37th Meeting of the Cognitive Science Society.

Weiskrantz L. Consciousness Lost and Found: A Neuropsychological Exploration. Oxford University Press, ; 1997

Werndl, Charlotte (2013). On choosing between deterministic and indeterministic models: underdetermination and indirect evidence. Synthese 190 (12):2243-2265.

Wesley, S. C (1984). Scientific Explanation and the Causal Structure of the World. Princeton University Press.

Wesley S. C. (1984). Scientific Explanation: Three Basic Conceptions. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1984:293 - 305.

Wilson, Deirdre & Carston, Robyn (2006). Metaphor, relevance and the 'emergent property' issue. Mind and Language 21 (3):404–433.

Wilson, Jessica M. (1999). How superduper does a physicalist supervenience need to be? Philosophical Quarterly 49 (194):33-52.

Wilson, Jessica M. (2005). Supervenience-based formulations of physicalism. Noûs 39 (3):426-459.

Wilson, Jessica M. (2006). On characterizing the physical. Philosophical Studies 131 (1):61-99.

Wilson, Jessica M. (2016). Grounding-Based Formulations of Physicalism. Topoi:1-18.

Witmer, D. Gene (2018). Physicality for Physicalists. Topoi 37 (3):457-472.

Woodward, Jim (2002). What is a mechanism? A counterfactual account. Proceedings of the Philosophy of Science Association 2002 (3):S366-S377.

Woodward, James (2003). Making Things Happen: A Theory of Causal Explanation. Oxford University Press.

Woodward, James (2004). Counterfactuals and causal explanation. International Studies in the Philosophy of Science 18 (1):41 – 72.

Woodward, James (2008). Causation and manipulability. Stanford Encyclopedia of Philosophy.

Woodward, James (2008). Mental causation and neural mechanisms. In Jakob Hohwy & Jesper

Kallestrup. (eds.), Being Reduced: New Essays on Reduction, Explanation, and Causation. Oxford University Press. pp. 218-262.

Woodward, James (2009). Agency and Interventionist Theories. In Helen Beebee, Christopher Hitchcock & Peter Menzies (eds.), The Oxford Handbook of Causation. Oxford University Press.

Woodward, James (2015). Interventionism and Causal Exclusion. Philosophy and Phenomenological Research 91 (2):303-347.

Worrall, John (2011). Underdetermination, realism and empirical equivalence. Synthese 180 (2):157 - 172.

Worrall, John (2014). Prediction and accommodation revisited. Studies in History and Philosophy of Science Part A 45:54-61.

Wright, Cory (2012). Mechanistic explanation without the ontic conception. European Journal of Philosophy of Science 2 (3):375-394.

Wright, Cory (2015). The ontic conception of scientific explanation. Studies in History and Philosophy of Science Part A 54:20-30.

Wu, Wayne (forthcoming). ‘The Neuroscience of Consciousness’. Stanford Encyclopedia of Philosophy.

Zhong, Lei (2011). Can Counterfactuals Solve the Exclusion Problem? Philosophy and Phenomenological Research 83 (1):129-147.

Yablo, Stephen (1992). Mental causation. Philosophical Review 101 (2):245-280.

Yetter-Chappell, Helen (forthcoming). Idealism Without God. In Tyron Goldschmidt & Kenny Pearce (eds.), Idealism: New Essays in Metaphysics. Oxford University Press.